uOttawa

Ethics in Social Autonomous Robots: Decision-Making, Transparency, and Trust

Fahad Alaieri

Thesis submitted to the University of Ottawa
in partial Fulfillment of the requirements for the
PhD in Electronic Business Technologies

Faculty of Engineering
University of Ottawa

Abstract

Autonomous decision-making machines – ranging from autonomous vehicles to chatbots – are already able to make decisions that have ethical consequences. If these machines are eventually deployed on a large scale, members of society will have to be able to trust the decisions that are made by these machines. For these machines to be trustworthy, their decisions must be overseen by socially accepted ethical principles; moreover, these principles and their role in machine decision-making must be transparent and explainable: it must be possible to explain why machine decisions are made and such explanations require that the mechanisms involved for making them are transparent. Furthermore, manufacturing companies have a corporate social responsibility to design such robots in ways that make them not only safe but also trustworthy. Members of society will not trust a robot that works in mysterious, ambiguous, or inexplicable ways, particularly if this robot is required to make decisions based on ethical principles.

The current literature on embedding ethics in robots is sparse. This thesis aims to partially fill this gap in order to help different stakeholders (including policy makers, the robot industry, robots designers, and the general public) to understand the many dimensions of machine-executable ethics. To this end, I provide a framework for understanding the relationships among different stakeholders who legislate, create, deploy, and use robots and their reasons for requiring transparency and explanations. This framework aims to provide an account of the relationships between the transparency of the decision-making process in ethical robots, explanations for their behaviour, and the individual and social trust that results.

This thesis also presents a model that decomposes the stages of ethical decision-making into their elementary components with a view to enabling stakeholders to allocate the responsibility for such choices. In addition, I propose a model for transparency which demonstrates the importance of and relationships between disclosure, transparency, and explanation which are needed for societies to accept and trust robots.

One of the important stakeholders of robotics is the general public and, in addition to providing an analytical framework with which to conceptualize ethical decision-making, this thesis also performs an analysis of opinions drawn from hundreds of written comments posted on public forums concerning the behaviour of socially autonomous robots. This analysis provides insights into the layperson's responses to machines that make decisions and offers support for policy recommendations that should be considered by regulators in the future.

This thesis contributes to the area of ethics and governance of artificial intelligence.

Acknowledgments

Completion of this doctoral dissertation was possible with the support of several people. I would like to express my sincere gratitude to all of them. First and foremost, I want to thank my supervisor André Vellino. It has been my honour to be his first Ph.D. student. Since I first began working on the thesis, Professor Vellino has offered valuable guidance, scholarly advice, and consistent encouragement throughout my research work. This this was possible only because of his unconditional support. I appreciate his contributions of time and ideas, the joy and enthusiasm he has for his research were contagious and motivational for me, especially during times of challenge in the pursuit of my Ph.D.

Besides my advisor, I would like to thank the members of my thesis committee who have guided me through all these years, Professor Rocci Luppicini and Professor Umar Ruhi, for their insightful comments and encouragement, but also for the hard questions which incentivised me to widen my research from various perspectives.

Nobody has been more important to me in the pursuit of this project than the members of my family. I owe a lot to my parents, who have encouraged and helped me at every stage of my personal and academic life, and who have longed to see this achievement come true. They are my ultimate role models. Most importantly, I wish to thank my loving and supportive wife, Asma Aljutaili, who has supported me in every possible way to see the completion of this work; and my two wonderful children, Saleh and Yaser, who have provided unending inspiration.

Table of Contents

List of Figures

List of Tables

1.Introduction

“Some electronics experts, including Bill Gates, predict that within a few decades robots will be as ubiquitous as computers are now” (Kernaghan, 2014, p. 486). These experts anticipate that robots will be in every house, every office, and maybe even every pocket — just as smart phones are today. At present, robots are being created to help patients in hospitals; to assist people, young and old, in their homes; to deliver packages; to carry people; to help customers; to make products; to discover space and oceans, and to rescue. In this regard, the spread of robots offers an opportunity to create a better world. As explained in detail later in this chapter, new statistics show an increase in the sale and use of robots in different sectors. This growth in the robotics market has had a significant impact on society as people rely on robots to make certain decisions and perform certain tasks that have ethical consequences. At the same time, it is important to note that pursuing robotic advancements without considering their ethical dimension is of considerable concern, because “historical experience shows that highly intelligent agents without ethical qualities may easily turn out to be unscrupulous and destructive” (Dodig-Crnkovic & Çürüklü, 2012, p. 61). Deng (2015) argues that creating an ethical robot is one of the most difficult challenges in artificial intelligence. Therefore, while it is more difficult to build an ethical robot than a normal robot, designing social robots without considering ethics is potentially dangerous. As I will show below there is too much potential for causing harm if ethical considerations are ignored.

Currently most robots, such as industrial robots on automobile assembly lines, perform only the specific tasks that are programmed into their systems. However, some intelligent robots

already have the ability to make decisions based on their surrounding environment. Other types of robots exist only in science fiction novels. As Kenneth Kernaghan says,

> It is important to distinguish between robots already in effective operation, those likely to be available in the near future, those on the distant horizon, those that are not only far-off but also far-fetched, and those that are unduly expensive. (Kernaghan, 2014, p. 487)

For the purpose of this thesis, I consider existing robots and near-term future robots. At present some robots are capable of performing diverse tasks for different people in various situations. Some of these tasks require software systems that make decisions and that have ethical consequences. There are currently many such decision-making software systems, for example systems that make recommendations, filter job applicants, offer mortgages, and predict crime rates, none of which currently have an explicitly ethical decision-making component. Moreover, some ethical decisions are already difficult for humans, let alone machines. For example, the trolley problem (discussed in Chapter 3) illustrates some of the difficulties involved in making decisions based on their ethical consequences. In this ethical dilemma, many people come to different conclusions about the right courses of action and there is no reason to think that ethical choices made by machines should be any more definitive. In addition, there are numerous inputs, situational variations, and scenario complexities in the trolley problem that lead to further challenges in machine ethical decision-making. Machines, such as autonomous vehicles in an unavoidable crash situation, will face the same difficulties as humans in making ethical decisions.

In this thesis I focus on both robots and bots — sometimes referred to together as (ro)bots (Wallach, 2010) — that have the capacity to make autonomous ethical decisions. Some contemporary robots have the ability to see, hear, speak, sense, move, and perform diverse tasks for different people in various situations. Some of these tasks require machines to make ethical decisions or exhibit ethical behaviours. Thus, in this research I examine robots that need to make ethical decisions and perform ethical actions. In particular this research focuses on the role of transparency and explanations in making such machine-decisions trustworthy.

In general the study of robot ethics focuses on three components: 1) the ethics of people who create robots, 2) the ethics that exist inside of robots, and 3) the ethics of people who treat robots (Asaro, 2006). The principal focus of this thesis is the second component of the ethics that are embedded in robots. To clarify, I do not restrict my study to robots that have a body; I also include bots that are basically computing systems. I differentiate between these types of robots later in this chapter.

Many robotics issues are related to ethics or have ethical and social implications. I do not cover them all, nor do I find new solutions to the problem of which ethical framework a robot should adhere. Rather, I focus on a framework to understand the relationship between manufacturing social autonomous robots and their impacts on society. Other natural questions that arise include: which ethical principles should guide social robots to make ethical decisions and perform ethical actions, what is the ethical decision-making process, and what are transparency, explanation, and their impact on societal trust.

First, it is essential to determine whether we need social autonomous robots to make ethical decisions, and if so, what types of ethical principles should be embedded in the ethical robot?

Should robots just follow fixed ethical rules, or should they make decisions based on the consequences of their actions? Should robots make ethical decisions based on their experiences that have been acquired by machine learning? Whose ethics/algorithms should be embedded in robots? Who is responsible for a robot actions? Should people know why and how robots make certain decisions? Is there an available explanation for any given robot's behaviours?

1.1. Increasing Demand for Robots

Many thinkers are concerned about the ethics involved in designing robots that are used in manufacturing, workplaces, homes, battlefields, roads, hospitals, and care centres. For instance, we are all concerned about the power of robots and their effects on employment, privacy, human dignity, social life, and even our lives. Up until now, there has been an absence of ethics in the design of robots. At the same time, there has been an increase in robot sales around the world. Different types of robots have been commercialized recently such as drones, autonomous vehicles, decision-making systems, recommender systems, chatbots, social robots, and healthcare robots. These robots have varying degrees of autonomy and ability to make decisions based on their calculations, and some of these decisions have ethical consequences. Notably, some robots have been involved in performing actions that are biased against groups of people because of the lack of ethics in their design -e.g. (Angwin, Larso, Mattu, & Kirchner, 2016; Bass, 2016).

To understand the concerns about the decision-making robots, we must consider some facts about the robotics industry. According to World Robotics 2017, a report produced by the International Federation of Robotics (IFR), there has been a noticeable increase in robot sales around the world in recent years. There has been a 16% increase in sold units of manufacturing

robots (e.g. industrial robots), a 24% increase in sold units of professional robots (e.g. medical robots), a 24% increase in service robots (e.g. household robots), and a 22% increase in entertainment robots (IFR, 2017a; 2017b). Using these statistics from the IFR report, the most important facts are summarized in Table 1. The two main drivers leading to the growth in sales and units of professional robots are the automotive industry and the electrical/electronics industry. First, almost all automotive manufacturers in the world use robots in their production processes. Therefore, the automotive industry is the most valuable consumer of professional robots. Second, electronics manufacturers purchase professional robots to produce TVs, communication devices, and medical equipment. Based on IFR estimates, between 2018 and 2020 there will be even more growth in robot sales. Of the various reasons for these projected increases I highlight the following: 1) ease of use gives more opportunities for all industries — including small organizations — to use robots; 2) competition within the robotics industry has led companies to focus more on robot research and development; 3) the growing robot market needs more players to meet the demand; 4) electronic manufacturers want inexpensive robots with short lifecycles for simple assembly tasks; and 5) the use of robots in manufacturing has resulted in clear improvements in work quality and work environment safety.

These trends are also apparent with bots, such as chatbots, personal assistant applications, and recommender systems which have manifested a corresponding increase in their development and use. For example, Facebook Messenger has more than 100,000 monthly active chatbots[1] which transfer two billion messages between customers and businesses each month. In

[1] Facebook. (2017, April). Messenger Bots for Business & Developers. Retrieved November 15, 2017 from https://messenger.fb.com/

addition, millions of smart phones and personal computers have been equipped with personal assistant applications such as Siri in iOS devices, Bixby in Samsung devices, and Cortana in Microsoft devices. Additionally, some companies now sell personal assistant devices such as Amazon Echo and Google Home. Finally, many websites are now equipped with recommender systems that help customers or visitors to buy products, read news, watch videos, read posts, or follow people, such as Amazon, eBay, Netflix, Google, Facebook, YouTube, and Twitter. All these bots and devices have some degree of autonomy and make some decisions on behalf of their owners that have an ethical dimension.

Given these clear growth trends in the robotics and bot industries, the environmental, social, ethical, and economic concerns that they raise should be carefully addressed. In any case, robotics companies which manufacture all those types of robots and bots, including autonomous vehicles, have a social responsibility towards society to create robots and bots that are safe, transparent, explainable, trustable, and reliable. For example, The Uber accidents[25] illustrate that these issues matter to the corporate bottom line. Note that there are no statistics about how many fully autonomous vehicles there are which are still in their development phase and have not yet been deployed.

Table 1 Facts and Estimations of Robot Sales

Robot Type	Example	Facts (2016)	Forecasts (2017-2020)
Industrial Robots	Manufacturing robots	- Production increased by 16% to 294,312 units - Sales value increased by 18% to US $13.1 billion - Worldwide stock increased by 12% to approximately 1.8 million units	- Increased by 15% per year to 520,900 units in 2020 - Between 2017 and 2020, approximately 1.7 million robots will be deployed - By the end of 2020, worldwide stock will be an estimated 3,053,000 units
Service Robots	Professional robots e.g. medical robots	- Increased by 24% to 59,706 units - Sales value increased by 2% to US $4.7 billion	- Between 2018 and 2020, almost 397,000 units will be sold - Sales value of US $19 billion
	Service robots e.g. personal and domestic robots	- Increased by 24% to 6.7 million units - Sales value increased by 15% to US $2.6 billion	- Between 2018 and 2020, approximately 32.4 million units will be sold - Sales value of US $11.3 billion
	Entertainment robots e.g. toy and education robots	- Increased by 22% to 2.1 million units - Sales US $0.98 billion	- Between 2018 and 2020, approximately 10.5 million units will be sold - Sales value of US $7.5 billion

The environmental, social, ethical, and economic concerns that bots and robots raise is front and center in the design of robots and the design of robots that make decisions ethically is becoming increasingly important.

1.2. Research Questions

If we allow robots to make decisions that take ethics into consideration, we need to ask:

1- In what situations and why might it be inadvisable to enable robots to make ethical decisions?

2- If humans are to trust robots to make ethical decisions, what are the conditions under which this trust is justified?

3- If decisions by robots are made using ethical principles, what is the role of explanations in establishing trust and allocating responsibility?

4- How do ethical decision-making algorithms need to be constrained in order to make them acceptable to the public?

These are the research questions that this thesis addresses.

1.3. Research Objectives

The main objectives of this study are to:

1- Identify the key issues for designing ethical robots;
2- Provide a framework that explains the relationships between robot manufacturers and society, and the role of regulators in making robots trustworthy and acceptable;
3- Help different stakeholders to provide a model for ethical decision-making in robots that enables a variety of stakeholders to understand the importance of ethics in making decisions about assigning roles and responsibilities; and
4- Illustrate the importance of transparency in the robotic industry to gain society's trust.

At the conclusion of this thesis I propose normative criteria that will assist different stakeholders such as governments, regulators, lawyers, manufacturers, consumers, and owners in deciding to whom responsibility should be assigned when robot actions cause harm.

1.4. Problem Statement

There are many concerns about enabling robots to make ethical decisions — the concerns about ethical decision-making discussed in this thesis are only a few among many that the general public have about robots generally. Mentioning some of these other concerns here enables us to put ours into context.

The first commonly expressed fear is that robots can replace workers in some factories. As robots become cheaper, more capable, more autonomous, and more intelligent, their role as replacements for human labour causes concern for a workforce that feels threatened by them.

The second fear is that the use of autonomous robots by the military "would increase danger to civilians" (HRW, 2012) because they would have the ability to kill without evaluating or calculating the ethical consequences of their actions. In response to this concern, many organizations around the world are working to ban dangerous robots such as killer robots[2] and autonomous weapons[3]. While this thesis does not directly address the question of autonomous military robots, the framework and models in this thesis could be applied to those types of robots as well.

The third fear is my focus in this thesis. Given that autonomous robots will have a greater ability to perform more tasks and make decisions with ethical consequences, even if an autonomous robot's behaviour can be controlled by algorithms that embody ethical principles, what kind of principles should they be? And who should be responsible for robots' critical mistakes?

1.5. Motivations

There are several motivations that led me to focus on this problem. They are:

1- Machines and systems (robots and bots) already have the ability to make decisions on behalf of people. Some of these systems are considered "intelligent" because their

[2] https://www.stopkillerrobots.org/
[3] https://autonomousweapons.org/

decision-making capabilities are enabled by artificial intelligence techniques such as machine learning.

2- Such machines and systems will have a greater ability in the future to perform more tasks for more people.
3- There is an absence of ethics in the design of these artifacts. Ethics is needed here to govern and control robots and bots' decisions which affect both individual users and society as whole.
4- There is a strong incentives among robotic companies to test and deploy their products in real-world situations without including ethical considerations in their design.
5- Absence of ethics and the presence of algorithmic bias has already led to artificial intelligence failures such as Microsoft's catastrophic chatbot Tay (Bass, 2016).

Take together these considerations point to the importance of the research questions listed in section 1.2.

1.6. Why do we Need Robot Ethics?

The emergence and spread of autonomous systems and machines have led to many concerns due to their impacts on societies. These concerns have spawned the emergence of organizations devoted to the issue of robot ethics. For example, the Berkman Klein Center at Harvard University and the Media Lab at Massachusetts Institute of Technology have collaborated to form the Ethics and Governance of Artificial Intelligence Fund[4]. A similar initiative called The Governance of AI Program has been developed by Oxford University's Future of Humanity

[4] https://cyber.harvard.edu/research/ai

Institute in collaboration with Yale University[5]. Therefore, it is important to formulate guidelines to control such technologies so that they remain human-centric and aligned with ethical principles and social values (IEEE, 2017). Technology is now part of our lives as humans; moreover, it shapes our lives in terms of interpersonal interactions and interacting with and relying upon other technologies. Essentially, "it is important to understand the nature and consequences of this new technology on human-robot relationships" (Arkin, Ulam, & Wagner, 2012, p. 572). In fact, we are connected to the technologies that we make. At times we create a new technology to fill a gap in an industry and it becomes part of the way we do things such as robotic arms in automotive industry. We create technology to interact with each other, to solve problems, to enhance our lives, and to make decisions. Over time our values change in response to different influencers, including technology; therefore, new values need to be realigned with technology's innovations (IEEE, 2017). Advancing our knowledge of robot ethics is important to shed light on the issues and problems that are generated by autonomous robots and to identify the basic issues that require further study and research. The rapid developments in the robotics industry brings problems, risks, and unclear issues and we need to identify the source of these problems and understand how to deal with them: "In short, technology's embedment in its sociotechnical system must also be taken into consideration" (E. Fraedrich & Lenz, 2016, p. 623).

Tzafestas (2016a) presents four benefits of conducting research on robot ethics. Studying robot ethics helps to: 1) clarify ethical behaviour, which in turn enhances ethics research; 2) establish a foundation of ethics, which in turn helps make ethics computable; 3) determine new problems in existing ethical theories, which in turn leads to the development of better ethical

[5] https://www.fhi.ox.ac.uk/governance-ai-program/

theories; and 4) give ethicists more opportunity to enhance ethics by working with artificial intelligence rather than ethicists working on theory only.

Technologies in general, including robots, have societal impacts on society related to employment, automation, privacy, human interaction, social values, ethics, humanity, dignity, corporate responsibility, algorithm biases, and so on. Thus, we need a framework with to help us understand the relationships between robots, the algorithms that control them, how they are designed, their manufacturers, their users and regulators and legislators. Unfortunately, not all roboticists are aware of the ethical implications produced by designing and using robots, and not all roboticists are experts in ethics. Some robotics companies succumb to market forces and build and release their machines before they have been tested from an ethical point of view. The rush to implement and deploy robots without ethical limitations could affect society in multiple ways and such ethical implications have occurred many times recently. Examples discussed later in this dissertation show cases when robots have made decisions which have resulted in ethical consequences because of the lack of ethical guidance. Roboticists need to ask whether their products' actions have ethical consequences for society and whether their products ought to embed ethical principles that guide their behaviours? However, these questions and issues cannot be answered and addressed by roboticists only. Roboticists need support from other experts in different fields, as highlighted at the end of this section. Sullins (2015) argues that roboticists need to have more skills in multidisciplinary fields such as legal and ethical reasoning to make better robots.

According to Luppicini and Adell, "The relationship between ethics and technology is of seminal importance to society and raises questions that continue to challenge learned scholars from a variety of fields and academic backgrounds" (Luppicini & Adell, 2008, p. 2). New

technologies in some situations produce new problems that need to be addressed in new ways. Moor points out that "… revolutionary technology, generates many ethical problems. Sometimes the problems can be treated easily under extant ethical policies" (Moor, 2005, p. 115). However, some ethical problems require the application of ethics to solve them; therefore, Moor (2005) adds that we need to produce new laws and rules to deal with such new ethical problems. The use of technology generates very important insights about the role of technology in society (de Graaf, 2016). de Graaf adds that understanding human robot interaction is very important in order to comprehend the role of robots in our lives and their ethical impacts on society.

As a society, what do we need from robots? What is their role in our lives? Are they just tools, or are they moral agents? There are different types of robots based on different societal perspectives. Veruggio (2006) divides robots into four types: robots that are merely machines, robots that have ethical dimensions, robots that are moral agents, and robots that are a new evolutionary species. Steinert (2014) divides robots into four similar categories: robots as tools, robots as recipients of ethical behaviour, robots as moral agents, and robots as influencers in societies. In both classifications, I focus on the third and fourth categories of robots as moral agents and their impacts on societies. In this thesis, I discuss the ethics that are embedded in robots and their ethical impacts on societies. We can view ethics in two dimensions of easy and hard (M. Anderson & Anderson, 2007). As humans, we use ethics in our daily lives to make ethical decisions and this is the easy part. On the other hand, ethics are not an easy field and we are not all experts in it. Thus, ethics have become an important factor that must be considered by robots designers in each and every project (Sullins, 2015).

In the future, robots will have more capacity to make decisions and perform actions on behalf of humans. They will have the ability to do many tasks for us, such as deliver products, carry us, decide who to hire, decide which types of news we see, decide which type of products we buy, decide who is more likely to be a criminal, and so on. All of these tasks have ethical implications which must be examined and studied. Therefore, robot ethics are the main domain of this thesis, as I focus on ethics that are embedded in autonomous robots.

1.7. Thesis Outcomes (Contributions)

In this thesis I have offered a narrative literature review which covers various aspects of robot ethics, beginning with defining robots and robot ethics in general. I have explored the ethical theories that can be applied to robot ethics, the notion of ethical responsibility (including corporate social responsibility), how these concepts apply to the problems of autonomous robots including autonomous vehicles in ethical decision-making situations, and the issues that arise from machine learning and corporate social responsibility. From the literature review I have identified certain areas which need further research, especially with regard to the relationships between the concepts of transparency, explanation, and trust. For that I have developed:

1- A conceptual framework which illustrates the relationships between different components that include regulators, robotics corporations, robots, and society. The framework covers different areas in ethics and the governance of artificial intelligence to enable human-robot trust in societies. I call this the “trust framework”.

2- An ethical decision-making model which identifies the steps that an autonomous robot should take to make ethical decisions. This model is an element of the trust framework.

3- A content analysis of human discussions about the ethical dilemmas faced by autonomous vehicles. This is a unique work that provides evidence to confirm some of my ideas, such as the concept that transparency enables trust.

4- Consequently, I concluded that there is a need for transparency in the robotics industry because transparency affects society's trust in robots. I believe that this need can be understood effectively within a trust framework. Therefore, I have developed a transparency model which demonstrates the different stages that are needed to enable transparency in robots. These stages are disclosure, transparency, and explanation. This model is another element of the trust framework.

5- In the disclosure stage of the transparency model, I have found that there is a need to label robots for the purpose of enhancing transparency; therefore, I have developed a labelling scheme which contains different components involved in disclosing a robot's ethical specifications to potential users.

1.8. Thesis Structure

This document is organized in six chapters. In Chapter 2, I define the types of robots under discussion. In addition, I define the research scope which focuses on robots' implications in societies. In that chapter I propose a methodology for collecting data for this research based on a plan to acquire information from various sources such as publications, organizations/government reports, and public forums. Also, I consider how to answer the research questions. In addition, explained the research design and research methodology. In Chapter 3 I review the literature by looking in depth at some terms and the relationships between them. I define the word "robot" from different perspectives to understand what type of robot we are focusing on in this work. I define Artificial Intelligence (AI) and Artificial

Moral Agents (AMAs). In addition, I identify "robot ethics" a primary subject of this thesis. Also, I explain the types of ethical principles that could be applied in robotics to address dilemmas such as the "trolley problem", one of the best-known ethical dilemmas. Additionally, I define Value Sensitive Design (VSD). Next, I explain certain requirements for ethical robots, such as understanding and moral responsibility, and I present the associated arguments about these requirements. Then, I look in depth at autonomous vehicles from their levels of autonomy, to the trolley problem, to their societal issues. Since machine learning is an important factor in this thesis, I discuss this form of AI together with the concept of algorithmic bias. Finally, I define Corporate Social Responsibility (CSR), which is the business-related aspect of this thesis. Next, Chapter 4 I propose my contributions of the trust framework, the ethical decision-making model, the transparency model, and the labelling scheme. Chapter 5 is dedicated to public forum content analysis. In this chapter I explain data locations and collection, data organization and analysis, results, and discussion of result. In Chapter 6 I summarize the thesis and my contributions, and then explain how I have answered the research questions. Some difficulties and limitations are identified in this research and for robot ethics in general. In addition, I consider my work could be expanded upon in the future. Finally, I summarize my final thoughts about social autonomous robots.

2. Methodology

While this thesis aims primarily at devising a conceptual framework for understanding robot ethics, it also has practical implications both for the design of social autonomous robots and for their use and consumption. My aim is to devise a framework which will enable a broad range of stakeholders to understand, classify, and reason about decision-making devices in general and robots in particular, insofar as they exhibit ethical characteristics in their behaviours. This framework will enable technologists who are designing decision-making systems to better specify the technical requirements for how these systems need to operate when their decisions have an ethical dimension. Also, this framework will give technologists a better idea about how users can trust autonomous robots to make such decisions. For example, with such a framework, regulators could request robotics companies to label each product for the purpose of conveying useful information for users such as autonomy level, capabilities, and the ethical principles they obey. I give an example of such a labeling scheme in 4.1.4.1. To develop the framework, I examined the decision-making process in depth, which will lead me to devise an ethical decision-making model for ethical robots.

2.1. Research Scope

Robots and bots that make decisions autonomously are only one category among dozens of types of robots that perform different tasks and are used in different sectors. Other types of robots, which are not discussed in detail in this thesis, include military killer robots, medical robots, industrial arm robots, and human-dependent robots. Some researchers distinguish between dependent robots and independent robots. Bruce Clough, for instance, calls them automatic machines and autonomous machines (Clough, 2002). On the one hand, an automatic

machine performs tasks that are pre-programmed into its system. On the other hand, an autonomous machine performs actions "free of outside influence". The same distinction could be made about automatic and autonomous bots. Thus, autonomous robots make decisions and perform actions based on their sensing of and computing about the environment that surrounds them. Another factor used to select the robots under discussion in this thesis is the presence of interactions with humans. This factor has been met, for example, in social robots and recommender systems. Thus, autonomy and socialness are the two factors used to select the robots that are involved in this study.

The two main foci of this research are on robots that make decisions that have ethical consequences and on the social implications of the decisions that these robots make. Positive or negative implications would affect the acceptance and trust of robots by people and organisations in society. Robot owners, robot users, and people who live with or are affected by robots will rely on robots only if they trust them. If there is no trust, society will reject these machines and possibly resist their introduction in their lives. While several components enter into the conceptual framework developed in this thesis, there are also many elements of robot ethics that are beyond the scope of this work.

Figure 1 shows the main relationship between five principal conceptual groupings that I have labeled Stakeholders, Computing, Ethics, CSR, and Sustainability. The framework I develop later in this thesis (in section 4.1) analyses the relationships among these groupings or blocks as well as some of their internal structure. Specifically, I focus in greater detail on two important blocks, namely Ethics and CSR, and develop a model for each of them. In the Ethics block I developed the decision-making model (in section 4.1.3) and in the CSR block I

developed the transparency model (in section 4.1.4.1), which is supported by the content analysis of on-line discussion groups, described in 5.

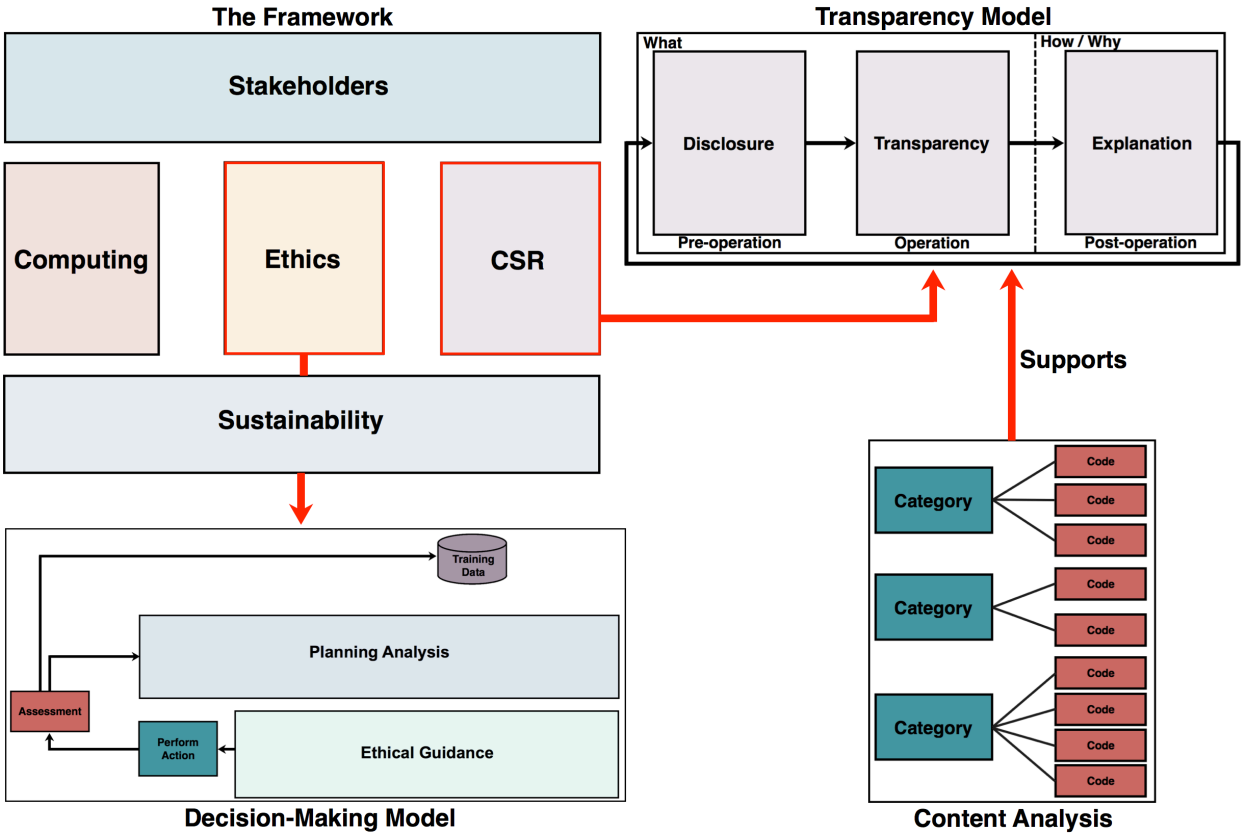

Figure 1 Research Scope

2.2. Research Design

To attain my research objectives, I first asked open-ended research questions that enabled me to explore the area of robot ethics and identify some of the conceptual problems that need to be solved. According to John Creswell (2014) this is Problem-Cantered research which belongs to a Pragmatic philosophical view. There are many benefits to using Pragmatic philosophy; for example, the researcher has the freedom to choose the method, techniques, and procedures that are necessary to meet the research needs and the researcher can use mixed methods (qualitative and quantitative) to collect and analyze data.

The research questions focus on the what, how and why of society's acceptance and trust in autonomous robots. According to Bhattacherjee (2012), there are three types of scientific research, exploratory, descriptive, and explanatory. The goals of conducting exploratory research are to determine the domain of the problem, to discover the core ideas about a phenomenon, and to examine the benefits of doing research that studies this phenomenon. The descriptive method aims to make detailed observation about the phenomenon. Therefore, in chapter 1 I first conducted exploratory research to identify the main problems in the literature concerning autonomous decision-making machines and the question of trust. Then, I conducted a narrative literature review (chapter 3) as a descriptive method, which led me to identify the core problems that are related to ethical decision-making and societal trust.

To clarify, in the descriptive methodology I collected documentary evidence from a variety of secondary sources about ethical issues in robotics. In addition to the published journal literature, I used documentary sources from reports in robotics labs, companies, partnerships, workshops as well as government policy documents. Each source addresses and focuses on one or more aspect of robot ethics. For example, the tenets of the (Partnership on AI, 2016) identify the need for openness and transparency in decision-making machines, as well as the role of responsibility, trustworthiness, and explanation. Some of the documentary sources focus on different aspects of machine decision-making, while others introduce some policies, regulations, and recommendations. I also apply the theoretical framework proposed in this thesis to a range of situations discussed in the literature in which I found discussions about decision-making robots. The issues raised in these cases focus on the ethical issues faced by the training methods in deep learning, algorithm biases, and the ethical theories used in decision-making.

From my survey of the literature, I found that there are some gaps that need to be filled. Firstly, there is a conspicuous absence of ethics in the design of autonomous robots: none of them have yet been designed with ethical decision-making capabilities. Yet some robots have already failed us by exhibiting unethical behaviour. Thus, I examined the role that ethics could play in governing the decision making in autonomous robots. The result of this analytic investigation led me to develop a decision-making model which shows how robots make decisions in general and ethical decisions in particular.

Secondly, I concluded from the literature review that the absence of ethics could affect robots' acceptance and trust by society, which, in turn, led me to enquire about what needed to be done to enhance societal trust in autonomous robots. I therefore developed a trust framework that shows the role that regulators play in controlling robot manufacturers and protecting societies, as well as the role that ethics plays in controlling robots' actions. This trust framework also shows the role that transparency of and explanation in the workings of these machines play in enabling trust in society.

Finally, I used this framework to "connect the dots" with a hypothesis that would explain why society would or would not accept or trust autonomous robots: my hypothesis was that the primary reason that ethical decision-making robots would not be socially acceptable was if they worked in non-transparent or non-explainable way. To determine when and how society would accept and trust autonomous robots such as autonomous vehicles, I conducted a content analysis of discussions in public forums where ordinary people share their opinions about autonomous robots. From that analysis I came up with a transparency model that illustrates the relationship between information disclosure, transparency, and explanations.

To substantiate my claims about the impact of the framework, I draw on the existing theoretical literature in philosophy, technoethics, robot ethics, and decision-making algorithms. In the literature review, I mention a variety of cases that contain decision-making machines and robots that perform actions which have ethical implications. These are discussed from different social and technical perspectives.

2.3. Research Process

As mentioned in the previous section, there are gaps in the literature that arise from the fact that there is an absence of ethics in both existing autonomous robots and in their design. These two factors are the cause of some of the failures in the decision-making process, in addition to the other factors that are discussed in chapter 3. This gap analysis led me to design a decision-making model that would take into consideration the ethics that govern robots' behaviours. This model was built on the Sense-Think-Act model (Beer, Fisk, & Rogers, 2014; Parasuraman, Sheridan, & Wickens, 2000; Vanderelst & Winfield, 2016) and it describes how to enable machines to make autonomous ethical choices. Following that, I designed the Trust Framework, which connects together the different factors (regulators, robots, computing, ethics, corporate social responsibility, and sustainability) that relates them to robot manufacturing companies and society. In an effort to understand the role of transparency and explanation for enabling trust, I conducted the content analysis to refine my Transparency Model (Figure 9), which describes the relationships between disclosure, transparency and explanation. I had arrived at several components of the Transparency Model from my gap analysis of the literature review and the function of the content analysis was to corroborate some aspects of this model.

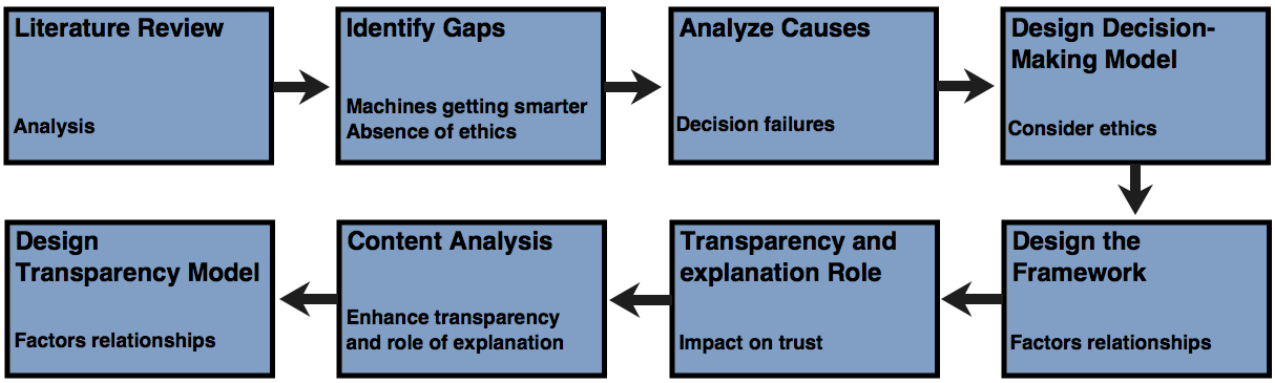

Figure 2 Research Process

2.4. Design Science Research Methodology

This research could be considered as Design Science Research (DSR). According to Hevner et al. (2010) DSR is a problem-solving paradigm in which the researcher answers questions about a problem by building an artifact as an end-goal of the research, which also serves to understand the problem. Artifacts and knowledge that are generated from the DSR methodology could be constructs, models, methods, instantiations and better design theories which contribute to the knowledge base in the area (Hevner & Chatterjee, 2010). Or the artifacts could be social innovations, specifications of technical or social resources, or any designed artifact that addresses a research problem with a solution. (Peffers, Tuunanen, Rothenberger, & Chatterjee, 2008). In general, according to Allen Lee (2001), Information Systems (IS) research investigates the phenomena that appears when technological systems and social systems interact. Hence, this method aligns well with the study of robot ethics as I discussed earlier in the section 1.6.

2.4.1. DSRM Process

According to (Peffers et al., 2008) the Design Science Research Methodology (DSRM) process includes six sequential steps which are: problem identification and motivation, definition of the objectives for a solution, design and development, demonstration, evaluation, and communication. Although, the steps are normally sequential, a researcher could start from any step and then move to other steps depending on the nature of the research idea. For this thesis, for example, I started at step 1 (problem identification and motivation) since my study is problem-centered research. The model in Figure 3, originally by Ken Peffers et al., has been adapted to explain the sequence of steps undertaken for this thesis.

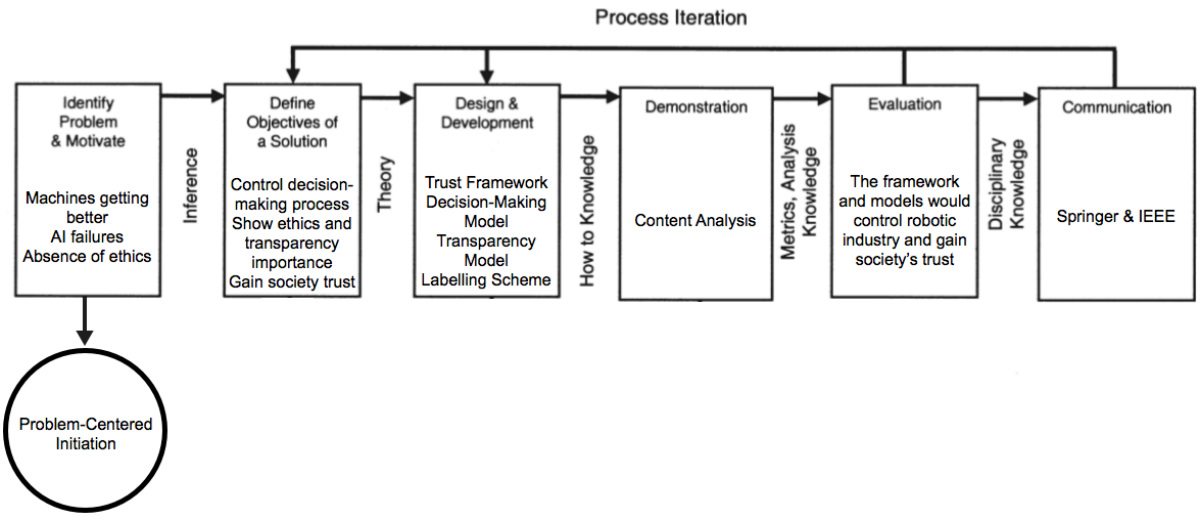

Figure 3 DSRM Process Model for this Research

Here is the explanation of each activity undertaken in this thesis that corresponds to this process model:

1- Identify Problem and Motivation: I noticed from the literature review that machines and systems getting better at making complex decisions; in addition, there have been a

number of artificial intelligence failures that resulted from the absence of ethics in the decision-making process of those systems.

2- Define Objectives of a Solution: according to Peffers et al. (2008) "Identified problems do not necessarily translate directly into objectives…" Thus, I needed to define the objectives of designing the artifact: a decision-making process that was able to make ethical decisions but which could also enable trust.

3- Design and Development: in this step I designed the trust framework, ethical decision-making model, transparency model, and labelling scheme as artifacts. The trust framework and the models are the descriptive outcomes from this thesis. The labelling scheme is a prescriptive outcome.

4- Demonstration: I conducted a content analysis of people's opinions in public forums to show the relevance and validity of elements of the trust framework and the models.

5- Evaluation: I evaluated part of the trust framework based on results of the content analysis. For example, the need for disclosure and transparency that were mentioned numerous times in the discussions.

6- Communication: some results of the work have been published in conference proceedings.

2.5. Design Theory

From the point of view of DSRM the outputs of this research, namely the trust framework, the decision-making model, the transparency model, and the labeling scheme can be considered Design Theory. In this research I explored the processes involved in the technology of decision making and the role of people and organizations that need or care about the trust that we place in these decisions using an analytic process in the literature review. I then determined the

requirements that are needed to move from the problem (the absence of ethics in decision making machines) to the outcome, which is a decision-making model that could be used to build trustworthy ethical robots. In the last step, I explored and defined the relationships between variables. For example, the impact of transparency and explanation on trust which affect companies' reputation. Hevner et al. (2010) would call these three steps Observation, Classification, and Association.

According to (Gregor, 2006), theories aim to describe, explain, and enhance our understanding of the world and sometimes yield predictions of what will happen and give us reasons to intervene and act in a certain way. In the purpose of this research, the theory that resulted from this work can be considered as theory for Design and Action because it offers an account of the processes that ethical decision-making machines perform as well as provide a model for prescriptive conclusions. According to (Gregor, 2006) theories have seven components that must be analyzed. Table 2 shows the components of the theory in this research.

Table 2 Research Theory Components

Theory Component	Definition
Means of representation	Words, Diagrams, Tables
Constructs	Stakeholders, requirements, responsibilities
Statements of relationship	Regulators must guide corporations to build ethical robot that is transparent to gain society's trust, which affects the reputation of corporations.
Scope	Framework that determines stakeholders' needs and responsibilities
Causal explanations	The combination of transparency and explanation is essential to enable trust in a society that uses and lives with robots
Testable propositions (hypotheses)	Transparency enable trust
Prescriptive statements	Regulators including governments and policy makers must guide robot companies to equip their robots with ethical modules to control their behaviours in order to protect societies from robot failures, to make embedded ethics transparent and to make robot actions explainable

2.6. Content Analysis Methodology

Chapter 5 is devoted the empirical analysis of text content in discussion forums. The content analysis was intended to provide evidence in support of the hypothesis that society's trust in robots would be enhanced by ensuring that there is transparency in robots' behaviours and explanations for those behaviours.

The methodology for this part of the thesis needs to be articulated clearly, as much for what it is as for what it is not. The analysis of the discussion forums is not the foundation for the theoretical parts of this thesis – rather it serves as limited evidence to support the Transparency Model as illustrated in Figure 1. Nevertheless, I had no preconceptions about what this content analysis would show and proceeded to use Grounded theory, which is a way to translate recorded data about social phenomenon to build a theory (Bhattacherjee (2012). To analyze the text data, I used an Open Coding technique to identify the main concepts that are hidden in the text.

The content analysis of this text is based on the methodological steps that are described in (Bhattacherjee, 2012) and (Creswell, 2014):

1- Select the sets of text that are related to the main problem (the making of ethical decisions by a specific kind of robots, namely autonomous vehicles).
2- Examine the text to have general sense about what the people saying and what are their main ideas.
3- Identifying and applying rules to divide the text (comments) into segments by summarizing each comment into simple words and dividing them into different segments based on their meanings.

4- Coding the text by using these summaries and using them to produce general description of the categories.
5- Analyzing the coded data to determine their frequencies, their context and how they are related to each other.
6- Representing the description and the categories by using some passages from the data to convey the findings. In this step I used figures and tables to represent the frequencies of the data.
7- Interpreting and translating the findings and the results.

According to Bhattacherjee there are several limitations to conducting content analysis. First, the coding process is limited to the content that is available in the text, so the codes emerge from the text only. Secondly, choosing text that published online only is exposed to sampling bias because in this case the researcher systemically ignored any opinion that was not published in these posting. However, one of the aims of content analysis in this research is to detect whether there were any patterns in the expression of thoughts. For example, I want to see how many people support the idea of enabling machines to make life-and-death decisions. Thus, I used statistical sampling methods to discover the frequency distribution of thought categories discovered in the text.

In conclusion, in this thesis I followed several different methods (exploratory, analytic, descriptive) to explore the literature and describe the problem, namely the absence of ethics in the decision-making process of autonomous robots. During my exploration and description of the problem I asked some questions and answered them by describing the roles of different contributing factors and articulating their relationships to the core problem. Understanding the role of these contributing factors roles and requirements of robot ethics led me to the proposal

that ethical robots need to be transparent, predictable, explainable for them to be trustworthy and to enhance people's acceptance.

3. Literature Review

In this chapter, I first review some of the frequently used concepts in the field of robot ethics, explain the general classification for robots in terms of their autonomy, and illustrate the differences between robots that are and are not powered by artificial intelligence. Then, I define robot ethics as a field and illustrate its importance with a discussion of the trolley problem, given its ubiquity in discussions of robot ethics, as well as discuss its new version in the context of autonomous vehicles: the tunnel problem. Additionally, I enumerate some of the ethical principles that are needed to guide robots to make decisions in such situations. I also discuss the problem of assigning responsibilities to robots for their mistakes. Autonomous vehicles are discussed in detail in this chapter since these vehicles are exposed to ethical dilemmas in decision-making. In addition, I briefly discuss machine learning and its relationship with ethical decision-making including algorithmic bias. Finally, the question of corporate social responsibility is also discussed for the purpose of clarifying the role that robot manufacturers play in making trusted robots.

3.1. Robot Classifications

Many types of robots are used in a broad range of fields such as business, healthcare, hospitality, manufacturing, and the military. Some robots are teleoperated and need to be controlled by humans, while others act autonomously (Kievit-Kylar, Schermerhorn, & Scheutz, 2012). However, the literature indicates that there are finer classifications for categorizing robot types. Niku (2010) introduced three classifications from three organizations: Japanese Industrial Robot Association (JIRA), Robotics Institute of America

(RIA), and Association Française de Robotique (AFR). Table 3 shows all of the robot classifications that have been approved by these three organizations.

Table 3 Classifications of Robots (Niku, 2010)

Organization		Class	Description
	JIRA	Class 1	Manual-Handling Device: operator actuated; multiple degrees of freedom
		Class 2	Fixed-Sequence Robot: uses a fixed, predetermined mechanism to perform successive stages of a task; difficult to modify
RIA		Class 3	Variable-Sequence Robot: easy-to-modify variant of class 2
		Class 4	Playback Robot: machine records motions of human operator and then replicates
		Class 5	Numerical Control Robot: a movement control program replaces the recorded motions of a class 4 robot
		Class 6	Intelligent Robot: has means to understand environment; ability to successfully complete a task despite changes in surroundings
AFR		Type A	Handling devices with manual control of telerobotics
		Type B	Automatic handling devices with predetermined cycles
		Type C	Programmable, servo-controlled robots with continuous or point-to-point trajectories
		Type D	Extension of Type C with the added capability of acquiring information from its environment

Of JIRA's six different classes of robots, RIA considers only four (3 through 6 in Table 3). AFR has its own classification. Our focus in this literature review is on the intelligent robot or the autonomous robot which are Class 6 and Type D in Table 3.

3.2. Robots and Artificial Intelligence

Not all robots are equipped with artificial intelligence and not all artificial intelligences are robots. Therefore, there are differences between these two terms and sometimes people use these terms in overlapping ways. To clarify, I define the terms robot, bot, artificial intelligence, and artificial moral agent.

3.2.1. What is a Robot?

There is no single, generally accepted definition of "robot". There are different types of robots that perform different tasks in different ways. Also, there are different institutions working in the field of robotics that define robots from different perspectives. Some scholars even hold different opinions about whether a robot is a physical machine, software, or both. I discuss these opinions at the end of this section.

In general, a robot is a machine that performs tasks autonomously or is teleoperated (NASA, 2009). However, this definition does not reflect the real meaning of robot because there are machines that perform some tasks independently but are not robots, such as vending machines. Some definitions describe the robot as a machine that performs tasks for humans, but that is also not an accurate definition because there are many machines that perform such tasks but are not robots, such as coffee makers. A more precise definition comes from the Robot Institute of America (RIA), which states that a robot "is a reprogrammable multifunctional manipulator designed to move material, parts, tools, or specialized devices through variable programmed motions for the performance of a variety of tasks" (Spong, Hutchinson, & Vidyasagar, 2004, p. 10). The Galileo Educational Network (GENA) has a similar definition of a robot: "a system that contains sensors, control systems, manipulators, power supplies and software all working together to perform a task" (GENA, 2003). According to these definitions, there are four components that a machine should have to be called a robot: (i) sensors (such as cameras to see or microphones to hear); (ii) software (which can be thought of as the robot's mind); (iii) manipulators (such as legs or wheels to move); and (iv) an electric power supply such as electricity, batteries, or solar panels. A complementary definition from the online *Oxford English Dictionary* states that a robot is "A machine capable of carrying out a complex series

of actions automatically, especially one programmable by a computer" (Oxford, n.d.). Those definitions do not mention or describe an autonomous robot that has the ability to make decisions without human intervention and act based on computations about its environment. However, Lin et al. introduce a concise but general definition that includes autonomous, teleoperated, biological, or software-driven "robot as an engineered machine that senses, thinks, and acts" (Lin, Abney, & Bekey, 2011, p. 943).

These more nuanced definitions are still not comprehensive, since some people also use the term "robot" to include non-physical machines such as software that executes information-centric tasks like decision-making systems, recommender systems, and chatbots. In an article for The Atlantic entitled "What is a Robot? The question is more complicated than it seems" (LaFrance, 2016), Adrienne LaFrance's tries to define the robot based on the history of robotics and the opinions of different scholars. LaFrance cites Kate Darling from the MIT Media Lab: robots must be viewed as embodied algorithms, but algorithms alone are bots, not robots. In contrast, LaFrance quotes Rob High, the chief technology officer of Watson at IBM: bots are entirely math and they do not have any embodiment. Allison Okamura, from the Stanford CHARM Lab, sides with Darling by stating that computers perform information tasks whereas robots perform physical tasks. Therefore, some roboticists believe that while robots must have a body, bots are computing systems without bodies.

Since both the embodied type of robot and the purely software-driven bots have many common elements and may be usefully discussed together, Wendell Wallach choses to refer to both of them as "(ro)bots". He says that "Agents within computer systems are often referred to as bots. Therefore the term (ro)bots seems a useful way to collectively refer to physical robots and software agents within computers and networks" (Wallach, 2010, p. 243). However, in this

thesis, the term "robots" refers to both robots and bots because the proposed framework, decision-making model, and transparency model could be applied to both.

3.2.2. Artificial Intelligence

Poole et al. prefer to refer to what is usually known as Artificial Intelligence by the term Computational Intelligence (CI), which the authors define as the study of designing an intelligent agent (Poole, Mackworth, & Goebel, 1997). The authors state that an Intelligent Agent is a system that performs activities intelligently based on its computations to perform an action, and that it is able to learn new experiences to improve its responses.

There are different definitions for the term "intelligence" and there is some debate about what constitutes intelligence. A widely accepted definition of Artificial Intelligence (AI) is a system that succeeds at imitating human behaviours, such as using language or playing chess (Turing, 1950). Kok et al. mention Turing's early definition and classify definitions of AI that exist in the research literature and in dictionaries into four categories: (i) systems that think like humans, (ii) systems that act like humans, (iii) systems that think rationally, and (iv) systems that act rationally (Kok, Boers, Kosters, van der Putten, & Poel, 2009). For example, a humanoid robot that has the ability to learn about its environment, learn about people's preferences, and recognize people's emotions is an intelligent robot. It is called an intelligent robot because it has a way of thinking that is similar to humans insofar as it explores the environment around it and recognizes and responds appropriately to people's needs. Thus, the definitions of AI and CI are similar. David et al. (1997) conclude that AI as a term leads to much confusion because of the word "artificial". Instead, the authors prefer the expression Synthetic Intelligence.

There is a diversity of applications of Artificial Intelligence in different areas and professions. In the now standard textbook entitled *The Principles of Artificial Intelligence,* Nilsson lists different principal applications for AI, including the controlling of robots and bots by AI agents. Such control systems incorporate many of the same techniques used in the other applications such as Natural Language Processing, Data Mining, Automated Reasoning, and Perception. AI-controlled robots need to sense objects in their environment, reason, make plans, solve constraint satisfaction problems, and make decisions (N. J. Nilsson, 2014).

3.2.3. Artificial Moral Agents

An Artificial Moral Agent (AMA) is an "artificial agent guided by [ethical] norms" (Nagenborg, 2007, p. 3). To understand this definition, we first need to define an Artificial Agent and clarify the term Agent. According to Nagenborg, an agent is a system that is designed to interact with humans. However, Franklin and Graesser first present many definitions for an agent from different sources and then introduce their own definition of an Autonomous Agent. For Franklin and Graesser (1997), an autonomous agent "is a system situated within and a part of an environment that senses that environment and acts on it, over time, in pursuit of its own agenda and so as to effect what it senses in the future" (Franklin & Graesser, 1997, p. 25). In this definition, the ability to sense the environment is required for a system to be called an agent. Talbot et al. argue that robots cannot be agents because they cannot act for a reason. To act for a reason requires an intention, and intention is the minimum requirement for agency (Talbot, Jenkins, & Purves, 2017). Therefore, a robot cannot act intentionally because it is not conscious and doesn't have intentions.

According to Nagenborg (2007), there are two types of AMAs: (Type1) agents are guided by specific rules. They act based on these rules without making any changes to these rules. (Type2) agents have the ability to create new rules based on other previously defined rules. The second type of agent is more intelligent than the first type because it has the capacity to make decisions autonomously based on its computations about its surrounding environment. AMAs are important from an ethical point of view because "AMAs should be able to make decisions that honour privacy, uphold shared ethical standards, protect civil rights and individual liberty, and further the welfare of others" (Allen, Wallach, & Smit, 2006, p. 13). From that perspective robots and bots as agents are required to be intelligent; more than being intelligent, they are required to be moral agents or at least have morals, principles, and values that guide their decision-making process.

3.2.4. Human Robot Interaction

Regardless of whether robots can be thought of as true (intentional) agents, it is undeniable that their actions have ethical consequences for humans. Therefore, the acceptance and trust of robots is gained based on the interactions between robots and humans. Attention to Human-Robot Interaction (HRI) has increased since robotic systems have developed more capabilities and interactions between humans and robots have became familiar (Clarkson & Arkin, 2007). Robot evolution into social environments produces legal and ethical challenges (Riek & Howard, 2014) and HRI is the field that studies how to design, understand, and evaluate robotic systems to be used by humans (Goodrich & Schultz, 2007). This interaction between robots and humans is based on their communications, which are in turn divided into two categories: *remote interaction*, when humans and robots are apart from each other; and *close interaction*,

when humans and robots are close to each other. Social robots usually perform close interactions with humans as peers or as companions.

According to Beer et al. (2014) there are five HRI attributes that designers must consider when designing a social robot: team, tasks, level of autonomy, information exchanged, and the adaption, learning, and training of people and robots. For instance, the robot's level of autonomy has significant importance in HRI because it affects the types of tasks that the robot can perform, the frequency of human interactions with the robot, and the reliability of the robot in society. Therefore, Beer et al. add that we should not ask "What can a robot do?" because this is not an important question. Rather, we must discover "What should a robot do, and to what extent?"

According to Scholtz and Bahrami (2003), the interactions between humans and robots are divided into five categories with humans functioning as robots' supervisors, operators, mechanics, teammates, or bystanders. Goodrich and Schultz (2007) added two categories: humans as information consumers, and robots as mentors for humans. It is crucial for HRI research to consider ethical, social, cultural, legal, and policy perspectives in order to evaluate the consequences of and benefits from HRI, such as trust, which is a key area of robotics (Kirkpatrick, Hahn, & Haufler, 2017).

HRI is important in the context of robot ethics because interacting with robots has ethical implications for humans. For instance, the appearance of robots in society — whether humanoid or not — could change how its members make decisions and perform actions; moreover, the responsibilities and roles of humans and robots in that society would be redistributed among them because people would rely on robots to shape their lives. Therefore,

explicit and implicit morals and values linked with technologies would enhance people's ethical practices (de Graaf, 2016). de Graaf adds that interactions with robots differ from interactions with other technologies because of the embodiment of robots which affects humans socially and emotionally; therefore, users deal with some robots as if they have emotions. In fact, much research has been conducted on the impact of ascribing emotions to machines. For example, research shows that building a human-like robot can make users lie to it to avoid hurting its feelings (Hamacher, Bianchi-Berthouze, Pipe, & Eder, 2016).

3.2.5. What is Robot Ethics?

According to the online *English Oxford Dictionary*, ethics by itself comprises a set of moral values that guide or control human activities (Oxford, n.d.). As humans, we acquire our ethical values from different sources such as religion, culture, the environment, and family as shown in Figure 4. Some of these ethical values are expressed in codes, regulations, and laws that were developed by certain people such as religious leaders or law-makers in a community. People in such communities use these ethical values in contexts to deal with each other, make decisions, and perform actions. These processes include reasoning, thinking, understanding, logic, intentions, preferences, beliefs, and desires.

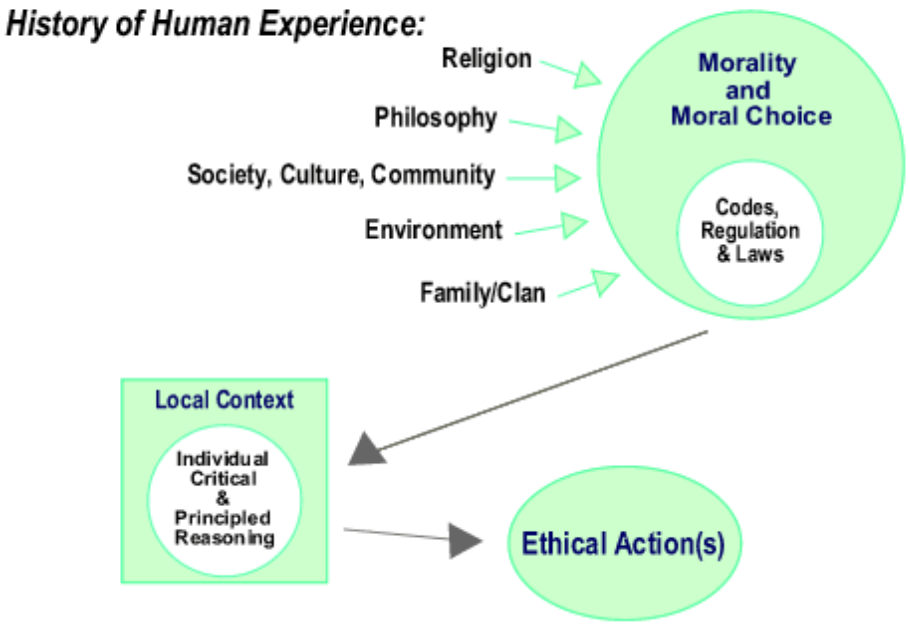

Figure 4 Origin of Ethics (Taft, 2000)

Therefore, robot ethics could be defined as a set of moral values that guide robots' activities. While this is the simplest way to define robot ethics, more details need to be added to this definition. To clarify, robot ethics form a type of applied ethics, which is "a branch of ethics which investigates the application of ethical theories in actual life" (Tzafestas, 2016b, p. 15). Applied ethics comprises different areas such as business ethics, media ethics, manufacturing ethics, technology ethics, and robot ethics. Consequently, robot ethics is a real-life application of ethics. In some aspects one application area covers another application area such as technology ethics and robot ethics.

The fields of technology ethics and robot ethics intersect. Technology ethics (sometimes called Technoethics) is an "interdisciplinary field concerned with all ethical aspects of technology within a society shaped by technology" (Luppicini, 2010b, p. 40). So, the field of Technoethics studies the social implications of technology. Thus, robot ethics form a branch of applied ethics that analyzes the issues that come along with creating and using robots (Tamburrini, 2009) and

also the need to protect people from robots' failures and their misuse. Different areas are involved in robot ethics such as robotics, computer science, artificial intelligence, philosophy, ethics, theology, biology, physiology, cognitive sciences, neurosciences, law, sociology, psychology, and industrial design (Veruggio, 2006). Robot ethics relies on different theories and solutions from those fields to address the ethical implications of robots on societies. Different related terms that suggest alternate meanings for "robot ethics" include machine ethics, machine morality, artificial ethics, and artificial morality (Allen et al., 2006). All of these terms focus on how to design machines and how these machines behave in moral way. As well, these terms suggest ways in which ethical principles can be incorporated into the design of autonomous robots.

According to Dodig-Crnkovic and Çürüklü (2012), robot ethics contains two main sub-fields in computer ethics: engineering ethics and machine ethics. Researchers in both fields maintain that engineers should design and create robots in ethical ways, and that machines should perform tasks based on ethical standards. Gianmarco Veruggio and Keith Abney as cited in Kernaghan (2014) divide robot ethics into three categories: 1) roboethics, 2) ethical behaviour of robot, and 3) ethical analysis and ethical decisions. The first category studies the ethics of designers, engineers, creators, and users of robots. The second category focuses on the ethical behaviours of robots. The third category — known as machine ethics or machine morality — focuses on autonomous ethical robots or Artificial Moral Agents (AMAs). In addition, some researchers differentiate between robot ethics and machine ethics, suggesting that machine ethics are not a type of robot ethics. Wallach says that robot ethics "addresses societal concerns in the deployment of robots", whereas machine ethics "considers the prospects for developing machines that are explicit moral reasoners" (Wallach, 2011, p. 196).

3.3. The Characteristics of Ethical Robots

Because robots interact with humans, safety is the number one priority in designing robots (Vanderelst & Winfield, 2016). However, safety is not the only important requirement of designing ethical robots. Designing an ethical robot is not easy. As Deng says, "Working out how to build ethical robots is one of the thorniest challenges in artificial intelligence" (Deng, 2015, p. 25). Even though this issue is difficult, it is possible for designers to control robots' behaviours by incorporating ethical principles into their decision-making systems.

According to Dodig-Crnkovic and Çürüklü (2012), "One of the essential characteristics of ethical behavior is moral responsibility" (p. 64). These authors claim that moral responsibility has two approaches: the classical approach and the pragmatic approach. The classical approach indicates that machines cannot be responsible; whereas the pragmatic approach of "artificial morality" posits that machines can be somehow responsible. Artificial morality the responsibility for ethical behaviours from creators and users onto the machines or robots (Allen, Smit, & Wallach, 2005). Accordingly, responsibility is one of the ethical robot's primary characteristics. The question of responsibility is discussed later in this chapter. First, there are several other characteristics that need more discussion.

A great number of researchers have written about ethical robots. Some discuss the rules that robots should obey to be ethical, while others present requirements that should be in a robot's system for that robot to be called an ethical robot. This discussion is vital because it is essential to first determine what type of ethics should be in the ethical robot. Do we need robots to make moral decisions in a utilitarian manner, one in which they evaluate the ethical outcomes of the possible choices they could make? Or should robots simply follow fixed ethical rules?

There is a difference between acting based on ethical rules and acting based on ethical consequences in different situations. For instance, having a robot that does not cheat, does not harm humans, or does not kill people is acting based on ethical rules regardless of the action's consequences. These actions are pre-programmed in the robot. The robot performs an action based on specific rules and it will perform the same action again every time it is in the same situation. For example, a vehicle that has been equipped with Advanced Driver Assistance Systems hits the brake once it detects an object in front of it because it has been programmed to do that. The vehicle does not have to calculate the consequences of the potential accident if it does not hit the brake, and it does not calculate the consequences of its actual action. On the other hand, a robot that does calculate and evaluate its actions before it performs them is acting based on a calculation of actions' consequences. For example, suppose that a fully autonomous vehicle in an unavoidable accident decides to turn and hit a tree instead of a child crossing the street. The vehicle could calculate and evaluate the consequences of its possible actions before performing the action and finds that hitting the tree causes less damage than hitting the child, especially since the vehicle has many safety features -e.g. airbags- for passengers and may conclude they will (likely) not get hurt.

Torrance (2007) discusses two examples of robots engaging in humans' moral lives. The first type is a robot that might act without ethical thoughts or moral sense while at the same time helping humans make moral decisions. The second type is a robot that acts somehow with a moral conscience. The second type also has an ability to recognize its environment, such as the fully autonomous vehicle; therefore, it could behave more responsibly than the first type of robot. Thus, there are two approaches to building an ethical robot, or to implementing ethics in an artificial machine so that it can make ethical decisions. Allen et al. (2005) describe these

top-down and bottom-up methods. A top-down strategy can be defined as installing or embedding the knowledge of what is moral and what is not in the machine. According to Lin et al., "one natural way to think about minimizing risk of harm from robots is to program them to obey our laws or follow a code of ethics" (Lin et al., 2011, p. 946). The robot then makes decisions based on these embedded ethical principles. Conversely, a bottom-up strategy can be described as a machine that has an ability to learn what is right and what is wrong from its environment. The machine with the bottom-up strategy has the ability to develop its own ethical system from experiences. This development can follow three courses. First, the robot could develop its system by the method of trial and error; second, the machine engineers could maintain the system and develop it with new moral behaviours; and third, the robot could adopt a continuous learning method which would allow the ethical robot to keep learning from its experiences and surroundings.

The top-down and bottom-up approaches have been discussed in a different way by Wallach (2009), who argue that there are two ways to implement ethics in robots: either the robot programmer installs rules in the robot's systems that mechanically lead to the desired outcomes, or the programmer builds an open-ended system that can collect information from its environment to predict the outcomes of its actions through a reasoning process. Similarly, Nagenborg (2007) divides Artificial Moral Agents (AMAs) into two categories — agents guided by specific rules and agents that have the ability to create new rules based on other rules that they have observed. Based on this consistent separation between the two approaches, it is clear that there are two types of ethical robots: one has built-in norms, and the other has a system that can learn from its environment. However, Moor (2006) divides the ethical robot into four types: namely the ethical-impact agent, the implicit ethical agent, the explicit ethical

agent, and the full ethical agent. The ethical-impact agent acts based on the consequences of an action. For example, an act would be performed if it leads to good consequences. The implicit ethical agent acts ethically based on rules that are programmed in its system. For example, an airplane's autopilot system is designed to take care of the airplane in the sky. Notably, this is an ethical behaviour because it sustains the lives of passengers. The explicit ethical agent acts based on calculations of decisions, actions, and their consequences with an ability to determine what must be done. The full ethical agent is almost as same as the explicit ethical agent but with the capacity of consciousness, intentionality, and free will, just like a human.

Can programmers install all morals or ethics in a robot by using the top-down method? Some ethical behaviours cannot be installed in a robot because they depend on other elements for which there are no good synthetic counterparts, such as emotions. Coeckelbergh (2010) says that emotions cannot be installed in a robot, but that they should and can be developed by the robot itself as a bottom-up process. In fact, when dealing with ethical robots, a hybrid strategy — or a method that combines the two approaches — might be the most effective (Allen et al., 2005). An ethical robot designed with a hybrid approach will have specific rules, but it will also have an ability to learn new rules and ethical behaviours from its environment.

When discussing the morality of robots, Wallach and Allen (2009) affirm that robot morality could be divided into operational morality, functional morality, and full morality. In operational morality, robot engineers design the robot's system and know all actions that could be performed by the robot, meaning that a robot's actions are pre-programmed, deterministic, and predictable. In functional morality, robot engineers design the robot's system but cannot predict the actions that could be performed by the robot, because the robot has the ability to

act and decide without direct control. This means that the robot's choices are determined by many input variables. Finally, in full morality, the robot has the ability to perform all actions autonomously. The robot's actions can be performed fully autonomously and its decision-making processes are not necessarily known. Figure 5, reproduced from Wallach and Allen (2009), shows the two dimensions of AMA development. There is a direct correlation between ethical sensitivity and autonomy, meaning that the ethical robot must have some degree of autonomy.

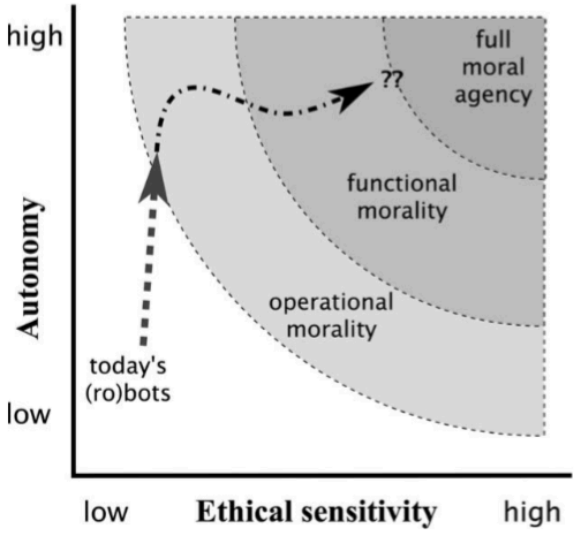

Figure 5 AMA Development Dimensions (Wallach & Allen, 2009)

3.4. Trolley Problem

The trolley problem is a well-known dilemma in ethics that serves as a useful test-case in discussions of robot ethics. Introduced by Philippa Foot in (1967), the main idea in this dilemma centres around whether it is ethical to sacrifice one person to save many. In this dilemma's general form, there is a runaway trolley on a track. Up ahead on the track there are five workers who will be killed by the trolley. There is a person standing next to a handle that could effectively switch the trolley to another track. However, this person notices that there is

one worker standing on the other track. If he pushes the handle, he will let the trolley kill one worker instead of five workers. Of these two options, what is the ethical decision? From a utilitarian point of view, killing one worker instead of five is the ethical decision. On the other hand, from a deontological point of view, the person next to the handle should not act to kill anybody.

Machines may will face the same kinds of difficulties in making decisions that humans face because numerous inputs, situations, and scenarios like the trolley problem could happen to autonomous robots and lead to challenges in designing their decision-making processes. For example, an autonomous vehicle in an unavoidable crash situation would need to explore its environment, identify the potential decisions, and analyze the consequences of these decisions so as to minimize harm. Finally, the autonomous vehicle must choose the most suitable action from among these alternatives. This decision could be made using a consequentialist analysis but may also be in harmony or in conflict with deontological rules such as "obey the law".

3.5. Ethical Laws and Theories

As stated by Rosalind W. Picard (1997), "The greater the freedom of a machine, the more it will need moral standards" (p.134). To explain the rules that ought to exist in an ethical robot or the requirements that should be in an ethical robot system, I must also discuss some ethical laws and ethical theories. The ethical decisions of autonomous machines should be made based on ethical principles and values that are built into robot systems, particularly autonomous robots that require no human input to make a choice. Since robots have an increasing ability to detect objects in their environment (via sensors) and compute a large number of possible

outcomes for the options that they consider, it is critical that the decision-making processes be governed by well-understood ethical principles.

As stated previously, ethical robots could have either specific ethical rules (laws) to control their actions, or they could have a learning system to develop their moral decisions. Installing laws in a robot to make it ethical is an approach that takes after Asimov's Laws, the first rules in machine morality. Asimov's three Laws of Robotics are:

> 1) A robot may not injure a human being or, through inaction, allow a human being to come to harm, 2) a robot must obey the orders given to it by human beings except where such orders would conflict with the First Law, and 3) a robot must protect its own existence as long as such protection does not conflict with the First or Second Laws. (Asimov, 1942, p. 27)

In a later novel, Asimov added a fourth law and called it the Zeroth Law: "4) A robot may not injure humanity or, through inaction, allow humanity to come to harm" (Asimov, 1985). The Asimov Laws describe characteristics that every single robot should have if they also possess autonomy and decision-making abilities. For Asimov, the narrative interest of these laws in his science fiction novels arises when there are apparent conflicts between these laws, such as when a robot is ordered by a human to do something that might harm another human. In a similar vein, Gert (1998) proposed ten rules that guide ethical robot actions: Do not kill, do not cause pain, do not disable, do not deprive of freedom, do not deprive of pleasure, do not deceive, keep your promises, do not cheat, obey the law, and do your duty. These types of specific rules that focus on ethical actions regardless of consequences are called deontological rules.

There are other capabilities that should also be available in a robot's system that contribute to its ability to be ethical. Gips (1991) introduces a list of requirements that an ethical robot should have: 1) a method to know its environment, 2) a system to calculate actions, 3) a method to predict the results, and 4) a way to calculate the consequences. In addition, Tzafestas (2016b) maintains that an ethical robot: 1) should have the ability to describe every situation in the world, 2) should be able to predict alternative actions, 3) should be able to predict the consequences of the current situation if an action is taken, and 4) should have the ability to calculate the goodness of a situation. Similarly, Sorell (2014) presents some requirements that must be included in the design of carebots (robots that take care of the elderly). The robot must be 1) autonomous, with the ability to set goals and find ways to achieve them; 2) independent, with the ability to apply goals without assistance; 3) enabled, with the ability to choose ways to achieve its goals; 4) socially knowledgeable, with the ability to contact family members and friends of the elderly person; 5) safe; and 6) private. Other researchers define some requirements in the context of agent responsibility (W. Loh & Loh, 2017). Loh and Loh (2017) define three requirements for an agent to be responsible:

1. Communication;
2. Autonomy;
    2.1. Know the consequences,
    2.2. Know the context,
    2.3. Personhood, and
    2.4. Know the impact range; and
3. Judgment;
    3.1. Cognitive abilities, and

3.2. Interpersonal abilities such as trust and reliability.

Each of these requirements could be met gradually in the evolution of ethical robots.

One over-riding consideration that determines the ethical characteristics of a robot is knowing which theories of ethics are used to determine what counts as an ethical or unethical action. Wallach and Allen (2009) admit, "the task of designing AMAs requires a serious look at ethical theory, which originates from a human-centered perspective" (p. 8). Sullins (2015) discusses several theories that have a relationship with robotics, including the theories of consequentialism or utilitarian (teleological) ethics, deontology or moral and ethical duties, virtue ethics, fairness and social justice, common goods, religious ethics, information ethics, and hybrid approaches. Each theory has its value for robot design, and robot designers must consider some of these theories in their projects. However, for the purposes of this thesis I focus on only two types of ethical theories: utilitarian ethics and deontological ethics.

According to utilitarian ethics — a form of consequentialism — the moral character of actions is judged by their consequences. John Stuart Mill (1864) in his book entitled *Utilitarianism* says, "Utility holds that actions are right as they tend to promote happiness and wrong as they tend to produce the reverse of happiness" (Mill, 1864). When applied to robot actions, a robot can be called an ethical robot if its actions lead to maximize human happiness. Winfield et al. implemented a prototype of an ethical robot that moves based on rules and by detecting its surroundings. This robot saves itself and other human lives by calculating the consequences of its actions (Winfield, Blum, & Liu, 2014). The robot was prototyped and tested in real life. In the experiment's scenario, there are two objects: a robot and a human. In the ground, there is a hole that is wide and deep. The robot has four options: stand still, move right, move ahead,

or move left. If it moves right, it will be safe and save a human life. If it moves left or stands still, it will be safe but the human will fall in the hole. Lastly, if it moves ahead, it will fall in the hole and the human will also fall in the hole. Figure 6 shows the robot's possible moves. This prototype illustrates that it is possible to implement a robot's actions based on utilitarian principles.

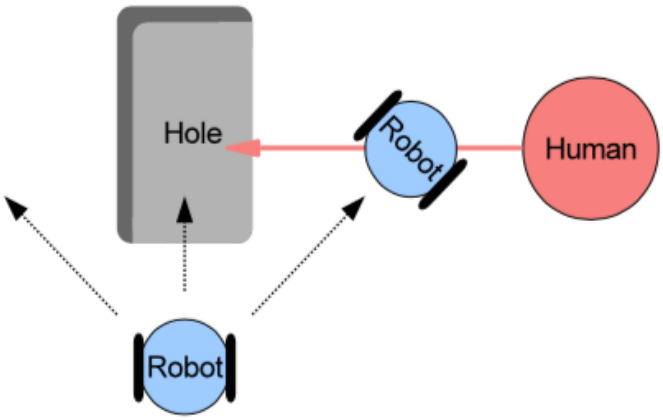

Figure 6 Scenario of Safety and Ethical Consequences (Winfield, Blum, & Liu, 2014)

However, Tzafestas presents five difficulties with utilitarianism: (i) it not easy to know who is affected by an action's outcome, (ii) different actions may lead to one outcome, (iii) the greatest pleasure cannot be calculated by a cost/benefit analysis, (iv) the greatest good for the greatest number of people could be harmful for some people; and (v) calculating what is right and what is wrong is a difficult and time-consuming process (Tzafestas, 2016b). Some of these difficulties fall under the category of "decision-making under uncertainty". Bhargava and Kim (2017) say that most of time decisions are taken in conditions of uncertainty. The authors add that although much research has been conducted on the subject of decision-making under uncertainty, studies have focused on empirical uncertainty. In contrast, moral uncertainty has

been given less attention, even though it is now very important — especially with the emergence of autonomous vehicles.

On the other hand, with deontological ethics both the wrongness and rightness of an action depend on the action itself, regardless of its consequences. For example, if an ethical robot is programmed to not kill, cheat, or harm, it must obey these rules regardless of the consequences. I would remark that: 1) consequentialism is often contrasted with deontology; 2) deontological rules resemble virtue-ethics insofar as it is possible in both systems for there to be a conflict between the rules (or virtues); and 3) a consequentialist and deontological analysis often coincide — for example some consequentialists adopt "rule utilitarianism". The ten rules Gert mentions above are deontological rules but they can also be viewed as utilitarian rules.

Tzafestas (2016) also introduced seven virtue-based ethical rules that could be used to guide an ethical robot: act with faith, act with hope, act with love, act prudently, act with fortitude, act temperately, and act justly (Tzafestas, 2016b). Tzafestas adds, "This is a multi-rule ethical system, and as in all multi-rule systems, it is possible to face a conflict between the rules" (p. 69). So, from the consequentialist point of view a robot's actions cannot be judged to be ethical or unethical until its consequences are known to at least a high degree of probability. Alternatively, a robot's actions can be considered ethical or unethical if they are determined to either be consistent with or inconsistent with some generally accepted deontological principles such as a principle of non-harming or a principle of honesty.

A more comprehensive method of deploying ethics in a machine is one which combines utilitarian rules and deontological principles (M. Anderson & Anderson, 2007). This approach could be necessary in the design of ethical robots in the future. An ethical robot designed as a

hybrid system could contain different rules from different ethical theories. Or, it could have specific rules from deontological theory (e.g. Asimov Laws) and guidance from a teleological theory (e.g. utilitarian principles). However, as I mentioned earlier, robots cannot (yet) be viewed as moral agents and so they cannot yet be blamed or praised for performing their actions. Therefore, Talbot et al. (2017) argue that because of uncertainty in the outcomes of future actions and human imperfection, building a robot to act like a consequentialist machine is right, even if consequentialism is false. Designers may design deontological robots as long as they are confident about the future decisions that robots will face. This confidence about future decisions may exist for a top-down robot in a closed and predictable environment that is known by the creators.

3.5.1. Value Sensitive Design

One approach to redressing this need is to embed ethical principles and values into the decision control system of robots by using Value Sensitive Design (VSD) to design a method that considers human values — for example, the social values of stakeholders — in the technology design process (Friedman, Kahn, Borning, & Huldtgren, 2013; Timmermans & Mittelstadt, 2014). According to Jeroen van den Hoven (2007), VSD "is a way of engaging ICT that aims at making moral values part of technological design, research and development" (van den Hoven, 2007, p. 67). Friedman et al. (2013) provide a list of human values such as human welfare, privacy, trust, courtesy, accountability, freedom from bias, and calmness. The application of Value Sensitive Design methods in designing such technologies would help mitigate the risks that users face in their interactions with systems that suffer from artificial intelligence bias and even outright failures.

To make ethically acceptable products that are trustworthy, designers first need to understand the values of end-users. These designers come to an understanding of users' values by talking to them, performing interviews, using questionnaires, conducting case studies, running focus groups, and also analyzing existing systems' feedback to determine their weaknesses and identify the design gaps. This step is especially important for those systems that make autonomous decisions that have ethical consequences. Given that robots that make decisions on behalf of users will be involved in people's daily lives such as in their connected vehicles, homes, and offices, the design of those machines and systems must take into account users' ethical principles and social values. For such robots to be accepted in society they will need to be evaluated against social norms and values (E. Fraedrich & Lenz, 2016). Yet is it possible to embed all social values in a robotic system? It may be possible to do that for social values that align with explicit ethical rules such as deontological rules, but what about implicit values? Some implicit social values (such as expressing false praise for the sake of social harmony) might be inconsistent with deontological principles (such as telling the truth) and yet yield desirable outcomes. As Isaac and Bridewell (2017) put it, "Ranking civility over truth may lead to false praise. Such false praise is technically deceitful…" (Isaac & Bridewell, 2017). They add that, as humans, we call false praise a white lie, but it is nevertheless "a lie". However, it may be acceptable to use false praise for a good reason, for example to be polite, and if we do this without immoral intentions. Thus, we can ask whether a robot ought to be polite and tell a white lie? For example, if children ask a social robot about the existence of Santa Claus, what should it say if it has the knowledge to tell the truth? If robots are to tell a white lie about Santa Claus, then how do we program them to distinguish acceptable white lies from unacceptable lies? Isaac and Bridewell (2017) claim that robots must not only recognize and respond to deceit but also be deceitful. Consequently, if lie-capable robots are taught not

only how to recognize a lie and how to lie, should they also be taught to what extent they should lie? But if a robot is capable of being deceptive in some circumstances in order to conform to social values, how could users trust it? How could users differentiate between white lies and ethical failures? Here, explanations might be useful. For example, if a robot can deceive and it deceives because one deontological rule (e.g. "civility") overrides another (e.g. "truth"), then it should be able to answer the question, "Why did you tell me I don't look fat even though I weigh 300 lbs?" with the explanation "Because not hurting your feelings is more important than the truth."

3.5.2. Understanding as a Characteristic

Natural Language Understanding could be a very important feature in an ethical robot that may be required to communicate with people who speak in a natural language. There is already some recognition of this in the growing presence of voice interfaces to computing systems (Siri, Cortana, Alexa, Google Home), even though these interfaces are far from having the ability to understand meaning. As early as 1980 Searle argued that a machine does not need to understand natural language to communicate with people. Searle's Chinese Room argument in his 1980 paper entitled "Minds, Brains, and Programs" (Searle, 1980) asks the reader to consider whether a person in a locked room who knows nothing about the Chinese language and yet is able to appropriately respond to sequences of Chinese characters with the help of a rule book is really "understanding" Chinese. If such interactions are successful, this means there is no need for true cognitive understanding on the part of the room in order to communicate effectively.

While Searle's argument is intended to support the claim that machines who operate with such mechanical rules are not "thinking", we can also use this argument to show that understanding is not necessary. Yet such processes take place in that way when computers communicate with humans: computer responses to natural-language input can be effective and appropriate without the presence of "artificial consciousness" or the ability to understand. Alternatively, some researchers argue with Searle (1980) by claiming that conversation is not only dependent on understanding, but that it is a part of it (Wallach & Allen, 2009). Some computing processes can be governed by rules, such as game-playing programs that follow explicit pattern-matching algorithms. But there can also be other processes used by open-ended systems that learn new patterns of play from experience. For instance, some computer games have rules and certain allowable moves (e.g. chess). The computer could learn all of the rules and then play the game based on these rules; however, it is difficult to install all rules and responses in a computer to make conversation in a natural language, because the game and natural language have different processes. (van Emden & Vellino, 2010).

3.6. Artificial Moral Agents and Moral Responsibility

Robots that make ethical decisions that have ethical consequences could be called artificial moral agents, but can they be held responsible for their actions? According to Noorman and Johnson (2014), "The discourse around responsibility and autonomous robots is rich and complex" (p.52). Therefore, there has been much debate about assigning responsibilities to artificial agents. Stahl (2006) states, "The traditional philosophical view is that non-humans cannot be subjects of responsibility" (p.209). However, other scholars argue that responsibilities can be assigned to artificial agents because they have become moral. As Asaro

(2008) notes, autonomous robots could have the ability to create their own rules and decision-making methods to perform tasks with both morality and autonomy (Asaro, 2008).

Some refuse the whole idea based on their position that artificial agents are only tools, which cannot be responsible for making decisions on their own. Bryson (2010) argues that we do not need artificial moral agents because robots are slaves to be used as tools which facilitate our lives and help us reach our goals faster. The implication is that robots are technological tools, equivalent to computers or cell phones, and when others are harmed by using these tools, then the responsibility lies with the user, not the tool. For other reasons, some scholars do not agree with assigning responsibilities to artificial agents because artificial agents do not have minds, free will, autonomy, or morality. Yet if robots have the ability to think, understand, make decisions, and act autonomously, then perhaps we can and should assign responsibilities to them. Robots now have the capacity to see, hear, talk, move, make decisions, calculate their actions' consequences, and act. That is why some researchers accept the idea of assigning responsibilities to robots as long as certain requirements exist in their systems such as ethical rules and the ability to make a decision. Indeed, much of the debate by researchers focuses on whether artificial agents can be responsible.

The following questions have been asked: Do we need to assign responsibilities to robots? Can we consider robots as moral agents? Do we need artificial moral agents that make moral decisions? Some researchers such as Talbot et al. (2017) say that robots cannot be agents. At this time, I believe that we do need to consider AMAs/robots that have the ability to make moral decisions based on ethical values because robots have already been given the capacity to make decisions. Moreover, some of these decisions affect human lives. Morality must be

programmed in robots' systems to guide their behaviours, and robots must be allowed to perform moral actions that lead to the greatest happiness.

3.6.1. Moral Responsibility

With the expanding diversity of autonomous agent types, "there is an expanding interest within Technoethics to assign moral responsibility to technology creators and creations" (Luppicini, 2010c, p. 62). W. Loh and J. Loh (2017) say that, in general, responsibility has five elements: 1) a subject that carries responsibility; 2) the role of the subject; 3) the subject is responsible toward an official receiver; 4) the receiver is the reason for the responsibility; and 5) criteria that determine what responsibilities the subject has towards the receiver. Stahl (2006) states that "responsibility can be external and internal … [that it] can have descriptive or normative aims…[and] have differing temporal directions" (p. 207). In other words, that responsibility can be ascribed to events in the future and to events in the past. Stahl's (2006) central point is that the question of who is responsible for what is socially contextual. One point to emphasize is that moral responsibility is not the same as legal responsibility, although there is an overlap between legal and moral obligations. Dr. A. Pablo Iannone from Central Connecticut State University in an interview with Luppicini (2010a) illustrates the distinction between moral responsibility and legal responsibility with an example. While promising to meet a friend for dinner at a certain time is a moral obligation, there is no legal obligation to keep this commitment to your friend because there is no law that forces people to meet friends for dinner. Hence, there is no legal liability if somebody fails to meet a friend. Stahl (2006) concurs with this distinction between legal responsibility and moral responsibility by stating, "While legal responsibility is institutionalised and attributed according to legal schemes, moral responsibility is not institutionalised" (p. 207).

Dodig-Crnkovic and Persson (2008) divide moral responsibility into two parts: causal responsibility and intention. Causal responsibility can be assigned to nonhumans, but only humans can have intention. However, Jarvik (2003) divides moral responsibility into three forms or types: causal responsibility, role responsibility, and liability responsibility. In the first form of causal responsibility, a person is responsible for everything that he or she has done before or will do in the future. In this form, the person is the cause of any action performed by him or her. In the second form, role responsibility, a person is obliged to perform a specific task, given his/her role in a certain area of society or community. In this form, a person's responsibility is his or her task, meaning that the task equals the responsibility. The final form, liability responsibility, refers to the question of who is "praised or blamed" for certain actions or outcomes. Before we praise or blame an actor for an action, we need to consider the various states of the actor, including their mental health and their intention as well as the consequences of the action, and the actor's influence on the action. Therefore, before we judge the moral responsibility of an actor and his or her action, we must analyze those conditions. Thus, responsibility has different forms and we need to specify which type of responsibility can or cannot be assigned to robots. Moreover, there are many different types of robots and not all types have the capabilities necessary to have any or all responsibilities ascribed to them. For example, an autonomous robot could have some moral responsibility because it is capable of making a decision based on its environment.

3.6.2. Discussion of Moral Responsibility in Machines

William Bechtel in his paper entitled "Attributing Responsibility to Computer Systems" presents some pre-existing arguments which emphasize that computers cannot be responsible agents (Bechtel, 1985). In his article entitled "A Question of Responsibility", Waldrop (1987)

notes that computers have become more intelligent; therefore, it is very important to consider what type of values should be installed in them. On the other hand, Waldrop (1987) adds "it seems unlikely that the question 'How much responsibility?' is ever going to have a simple answer" (p. 38), because the discussion about responsibility, authority, and control among humans has existed for thousands of years. This debate about how to assign responsibility for an action among has been going on for a long time and the recent improvements in artificially intelligent agents has perpetuated it. Stahl (2006) notes that while there has been a long standing debate about whether computers can be responsible, some conditions are required: agency and personhood. If computers fulfil these conditions, then they can be responsible; otherwise, they cannot.

While some researchers have attempted to prove that artificial agents cannot be responsible or cannot be moral others have focused on how to install rules in an artificial agent's system to guide its actions. In other words, some researchers have approved the idea of assigning responsibility to artificial agents and are now beyond the debate and at the point of developing ways to control artificial agents' actions by installing moral rules.

In the following subsections, I introduce arguments that are against assigning responsibility to robots, followed by a discussion of other arguments. Different authors reject the idea of responsible artificial agents for different reasons; for instance, they argue that robots are not moral, that robots cannot be moral, and that robots are not fully autonomous. Moreover, some argue that artificial agents cannot be responsible because they lack 'mens rea' - the ability to have a guilty mind. Since artificial agents do not have minds that are able to distinguish between rewards and punishments, they could make critical mistakes. The rejection of assigning responsibility to robots originates from various perspectives, itemized below.

3.6.2.1. Against Assigning Responsibilities to Robots

Proponents of this view argue that robots cannot be assigned responsibility because robots lack morality, personhood, autonomy, consciousness, mentality, intention, causality, and emotions (e.g. feelings of guilt). Stahl (2006) considers three arguments against artificial agents being assigned responsibility, namely: 1) computers do not have consciousness, so they do not have intentions; 2) computers lack fear, so they cannot be punished; and 3) computers lack free will. In the same context, others argue that artificially intelligent agents cannot be responsible because they do not have intention and they cannot be praised or blamed (Dodig-Crnkovic & Persson, 2008). Therefore, computers cannot be responsible because they do not have these human characteristics. Torrance (2007) posits that since artificial agents lack certain human characteristics or "properties of biological organisms", they are incapable of being full moral agents.

Some people reject the morality of robots by asserting that robots do not have minds or free will, and that robots are not autonomous. In addition, some argue that robots cannot be moral because they do not have a method for decision-making and they are pre-programmed. Naturally, as an extension of this argument, if robots are not autonomous then they cannot be moral. Selmer Bringsjord as cited in Sullins (2006) concludes that robots cannot become moral agents because they are not autonomous and they never will be, since they cannot perform any task that has not been programmed in their systems. However, it is important to note that Bringsjord's statement was made before the improvements and successes of deep learning artificial intelligence (Schmidhuber, 2015). Sullins (2008) states that artificial agents are neither moral agents nor legal agents because their actions are results of programs that interact with inputs and the environment, and they do not have minds equivalent to humans. Sullins

thinks there is a chain of conditions that is required for an artificial agent to be responsible. At the very least it must be able to make decisions as a function of a moral code and it must be able to make these decisions autonomously.

According to Dennis et al. (2015, p. 45), "Ensuring that autonomous systems work ethically is both complex and difficult". They believe that autonomy is one requirement for artificial agents to be moral. This means that robots cannot be moral while they are being controlled by users. Robert Sparrow also believes that if robots are not fully autonomous, then they are not morally responsible for their actions because "autonomy and moral responsibility go hand in hand" (Sparrow, 2007, p. 65). Sullins (2006) presents three requirements for a robot to be moral: autonomy, intentionality, and responsibility. In the first requirement, the agent should be independent and act without any control from either the user or from another agent. The robot must be fully autonomous to be considered a moral agent. In the second requirement, actions should be done by the agent with intention. While this does not require that the robot have full intentionality at the same level as humans, the robot should be clearly and evidently acting in a way that is "deliberate and calculated" (Sullins, 2006). In other words, we should be able to see that the robots know what they are doing. Finally, Sullins (2006) believes that "we can ascribe moral agency to a robot when the robot behaves in such a way that we can only make sense of that behaviour by assuming it has a responsibility to some other moral agent(s)" (Sullins, 2006, p. 28). In a slightly different argument, Johnson (2006) believes that computer systems can be moral entities but cannot be moral agents. Johnson claims that although computers cannot be moral agents by themselves, they are components of the moral agency of humans. In other words, designers and creators create computers to act in specific ways in particular environments at certain times to perform tasks for specific users.

Accordingly, computers are pre-programmed and they cannot be autonomous moral agents while they lack free will. In addition, considering computer systems as moral agents and ascribing responsibilities to them is hazardous because they can behave differently than humans and they can take actions that cannot be controlled by humans. In other research, Johnson (2014) calls this the 'Responsibility Gap' which means when these computers and artificial agents become more autonomous, no one can predict their decisions and no one will be responsible for their behaviours.

Artificial agents' free will is another main argument in this debate. Freedom of choice is a condition for an agent to be a moral agent. According to Stahl (2006), agents must be free and this is a "vitally important precondition" of being responsible. In addition, Bringsjord and Himma (2007) as cited in Sullins (2008) suggest that artificial agents cannot be moral because they do not have free will. Therefore, free will is an essential element that must be found in an agent's system if it to be considered a moral agent; only then can it also be considered a responsible agent. Tonkens (Tonkens, 2009) adds rationality to freedom as the two factors that must exist in an agent for it to be a moral agent. Yet does free will mean that to be moral, an agent must be able to act without constraints? What about the moral rules that humans obey? Do they rob humans of agency or responsibility? Moral agents must be guided by norms to control their behaviours. But that means agents should act based on rules, and such rules will not allow them to act freely. Thus, moral agents which have moral rules such as Asimov's Laws installed in their systems are not free, since "this would remove their free will" (Kopacek & Hersh, 2015, p. 76).

Wallach (2015) imagines that there might be intelligent agents in the future that have the ability to make complex decisions. Such agents could recognize their actions whether needed or not;

however; “holding the machine responsible for its actions is nonsense” (p.239). Although Wallach and Allen (2009) support the idea of moral robots, the authors are against giving robots all of the responsibility for their decisions and actions. They argue that machines act morally by having rules and values installed in their systems. Consequently, moral machines perform tasks in appropriate ways but only based on certain moral rules that are predetermined by humans. Therefore, humans have a role to play in taking responsibility for a machine’s actions.

3.6.2.2. Assigning Moral Responsibilities to Robots

The question itself of whether moral responsibility should be assigned to robots itself has an ethical dimension. According to Stahl (2006), it could be argued that “Ascribing responsibility to computers would… be desirable because of the good that comes from it and should thus be considered” (p. 209). Computers may have advantages over humans in terms of the accuracy and independence with which they make their decisions (M. Anderson & Anderson, 2007). Computers may have the ability to rapidly enumerate all actions that can possibly be performed in a certain situation and they may also be are able to calculate most of consequences that could result from each action. This accuracy could be achieved without direct control and without favouring themselves over humans. With these innate capacities they should be able to perform their social tasks (Stahl, 2006), both perfectly and accurately.

If we understand “responsibility” to be role-responsibility, then it is already true that robots are responsible in virtue of the role they play. For example, Japanese researchers are experimenting with robots which help police. These ‘urban surveillance robots’ are responsible for identifying criminals and detecting unusual behaviours (Royakkers & van Est, 2015). As another example, an ATM is responsible for giving money to customers based on a decision

that ATM makes. Also, the bank system is responsible for blocking a customer's credit card when it detects an unusual pattern. Therefore, computers are already responsible because of their ability to calculate, detect, inspect, and track.

Arguably, these types of computers are automatic agents, not autonomous agents, and therefore they are not themselves morally responsible. According to Clough (2002), an automatic machine is a machine that only performs specific tasks that are pre-programmed in its system. On the other hand, an autonomous machine is a machine that can perform activities "free of outside influence". Therefore, automatic computers cannot be responsible because they are not autonomous, and autonomy is one of the requirements for assigning responsibility. Another argument which has been raised about machine responsibility asks who is responsible if the machine makes a mistake, such as an ATM giving too much money to a customer. In this situation, who is liable or accountable? Do we blame the ATM, the ATM's manufacturer, the designer, the programmer, the bank, or the user (for example, if the user ends up keeping the money and not informing anyone of the error)?

At present, robots have the capacity to interact with their environment, make decisions, perform tasks, and calculate the consequences of their actions. In addition, robots learn from their experiences and they further develop their learning systems. These abilities lead to the conclusion that robots are morally responsible for tasks that lead to moral consequences (Dodig-Crnkovic & Çürüklü, 2012, p. 65). The capacity for robots to learn is a significant feature that encourages humans to assign responsibility to robots (Hellström, 2012). Therefore, a robot that learns from its experience and further develops its own system's capabilities can be afforded responsibility. Asaro (2008) predicts that in the future autonomous robots will obtain a greater ability to make their own moral rules, goals, and reasoning methods, and they

will thus be equipped to make moral decisions that fulfil the moral responsibility which has been assigned to them. I think that this prediction is based on an extrapolation from the current state of development of autonomous systems in the robot industry. For example, a number of companies such as Google, BMW, Mercedes, Audi, Volvo, and Delphi are currently developing autonomous vehicles that have the ability to drive without direct control by their drivers; moreover, some of these products have already been successfully road-tested. These types of cars will have the role-responsibility to carry people, including people who have disabilities, but also the moral responsibility that arises from their being able to make decisions autonomously.

Therefore, as I mentioned previously, certain conditions are needed for an artificial agent to be considered a responsible moral agent. Therefore, the installation of moral rules in agents' systems is needed to convert them into moral agents. Other researchers have introduced other moral rules, methods, and ethical theories that could be installed in a robot's system for it to act morally. For example, a robot/AMA that has an ethical dimension in its system, whether it is deontological or utilitarian, would be able to follow ethical rules or compute the least harmful course of action and to control its behaviours (S. L. Anderson, 2011). Building moral robots is possible; therefore, the morality conditions that are required for assigning responsibility can be achieved.

As I indicated above some thinkers believe a necessary condition for agents to be moral is free will. However, other researchers argue that free will, emotions, or even full consciousness are not required for an agent to be a moral agent. Floridi and Sanders (2004) support the idea that it does not matter whether robots act based on specific rules in a closed system or act autonomously in an open system, noting that computers can be moral artificial agents without

free will, emotions, or mental states. For example, animals could be considered moral agents even though they do not have free will, emotions, or mental states. As another example, children are considered moral agents even while lacking full consciousness and free will. Although they cannot be held responsible for some of their actions, they are still moral agents. Opponents of this idea may ask whether even we as adult humans have full free will and freedom over our actions. Some of us act based on tradition, beliefs, religion, culture, education, preferences, and social limitations. Therefore, although we do not have full freedom in our choices and we are not independent, still we are morally responsible agents. So, should robots be considered responsible moral agents only if they have free will and autonomy? Sullins (2006) argues that "robots may not have it, but we may not have it either" (p. 27).

3.6.3. Who is Responsible?

An essential question for society and legislators is if any harm is caused by our technological inventions, who is responsible? In the future, autonomous robots may have to shoulder full responsibility for their actions and outcomes (Malle, 2015; Noorman & Johnson, 2014). However, some researchers, including Johnson (2006), believe that it is dangerous to give robots full responsibility for their actions because they might act in ways that go beyond programmers' control. They might be called autonomous because they perform tasks without direct human control, but humans — be it the manufacturers, designers, programmers, or users — must take legal responsibility if anything goes wrong, particularly since robots are not yet legal persons. Hew (2014) claims that "[robots] will carry zero responsibility" for their actions, and that this responsibility should remain with humans (p.197). If users and manufacturers of autonomous robots are allowed to abrogate their responsibility for the harm they might inflict, this opens the door to a world in which the malicious hacking of such robots is responsibility-

free. Therefore, Wallach (2015) argues, people and corporations should be held responsible for all harm that is caused by technology. Kuflik (1999) agrees, concluding that the responsibility of robots' outcomes rests with the people who design them and who program their systems. To Kuflik (1999), the responsibility does not rest with the robots, because they are just programs running on machines. Responsibility for any action performed by a robot may have to be divided amongst different people such as the robot manufacturer and the user, and each group will hold part of this responsibility (Asaro, 2014; Tzafestas, 2016b).

However, some researchers believe that in some situations, robot users may hold all of the responsibility if they use their technologies in an intentionally negative way for the purpose of harming others (Crabb & Stern, 2012). Manufacturers argue that if a car owner turns on an autopilot system in an autonomous car and the car collides with another car or kills people, then the driver must take full responsibility for this outcome since the car was running in its fully autonomous mode and the owner made this choice. So, in this action or other actions that are performed by autonomous robots, there should be a responsible human agent in cases where something goes wrong. However, there are clear rules for allocating responsibility that have been produced and signed by many experts that link technology and responsibly. The formulation used by Miller (2011) in 'Moral Responsibility for Computing Artifacts: "The Rules"' is as follows:

> 1) People who design artifacts are responsible for these artifacts and for their decisions and actions;
>
> 2) The responsibility in the case of any harm is divided amongst those people who are involved;

3) The user of that artifact is responsible also;

4) "People who knowingly design, develop, deploy, or use a computing artifact can do so responsibly only when they make a reasonable effort to take into account the sociotechnical systems in which the artifact is embedded" (p. 58); and

5) The designer or creator of an artifact must not deceive users. People who design, develop, program, create, deploy, and use artificial agents are responsible for their agents according to their role in the action, decision, result, or harmful effect.

Consequently, these rules illustrate that humans still hold all the responsibility for their artifacts and property. Currently, robots are not responsible for their mistakes. Robotic corporations are responsible for their mistakes in robot design, and users are responsible for their mistakes in their use of robots. Policy makers and governments must inform and educate users about this responsibility while guiding robotic corporations to be reliable and dependable toward users.

3.7. Autonomous Vehicles

The emergence of autonomous vehicles (also called self-driving cars, automated vehicles, or driverless cars) produces different ethical challenges. Discussion and research about ethics in autonomous vehicles is essential and critical as many automobile corporations work hard to develop and test their autonomous vehicles. Moreover, other companies that specialize in electronics, learning, and transportation seek to develop and test their products that are related to or operate autonomous vehicles. For example, as of April 1, 2018 fifty-two corporations have obtained Autonomous Vehicle Testing Permits from the Department of Motor Vehicles

in the United States to test their products in California, including Mercedes Benz, Waymo, Tesla, Nissan, GM Cruise, BMW, Honda, Ford, NVIDIA, Udacity, Uber, Apple, and Samsung[6]. Some of these companies are working to develop their own vehicles that could be sold in the market for commercial use in the near future, while others are developing systems that control autonomous vehicles. In Canada, BlackBerry QNX received approval from the Ministry of Transportation of Ontario on November 28, 2016 to test autonomous vehicles on Ontario roads[7].

The ethical issues in autonomous vehicles are complex and, in some cases, unclear. Lin (2016) illustrates the importance of ethics in autonomous vehicles by explaining a scenario of unavoidable crash. In this scenario an autonomous vehicle needs to make a decision to swerve right and hit an eighty-year-old grandmother, swerve left and hit an eight-year-old girl, or continue straight and kill them both. The autonomous system in the vehicle must decide which option is the most ethical (or least unethical). Even if both actions are wrong, according to Bonnefon et al., (2015) "Some accidents … will be inevitable, because some situations will require AVs to choose the lesser of two evils" (p.1). Hitting the grandmother could be less immoral because she has lived a full life already while the girl is still innocent and young with a long life in front of her. In this situation and these circumstances, choosing between two lives to sacrifice one is not easy because the decision-making process relies on one factor to calculate and evaluate, namely the age. Consequently, both decisions are unethical because they are dependent on a type of discrimination that depends on age, which is morally on par with discrimination based on colour, race, gender, or religion. One component of the IEEE

[6] State of California. (2017, November 16). Testing of Autonomous Vehicles. Retrieved April 27, 2017 from https://www.dmv.ca.gov/portal/dmv/detail/vr/autonomous/testing

[7] QNX. (2017). Innovation New Ideas to Products. Retrieved November 30, 2017 from https://www.qnx.com/content/qnx/en/blackberry-qnx-autonomous-vehicle-innovation-centre.html

code of ethics teaches the following: "to treat fairly all persons and to not engage in acts of discrimination based on race, religion, gender, disability, age, national origin, sexual orientation, gender identity, or gender expression"[8]. As Gerdes and Thornton (2016) suggest, "a purely consequentialist approach using a single cost function to encode automated vehicle ethics seems infeasible" (p. 92). Therefore, what is the best action to take in that situation? Should the autonomous vehicle continue straight? The consequence of this decision would be more harmful than either of the first two.

Given that autonomous vehicles cannot yet distinguish human obstacles according to such features as age or gender, this is still a hypothetical problem. Yet we need to consider these ethical design issues in the development of autonomous vehicles because they are not easy to solve (Lin, 2016). Autonomous vehicles need to be programmed to respond to a variety of circumstances that might arise in unavoidable crashes, and programmed for how to crash (Nyholm & Smids, 2016), even though it is a very difficult challenge to design algorithms that could make such decisions (Bonnefon, Shariff, & Rahwan, 2016). Another challenge is knowing how to identify best consequentialist principles that estimate the costs and, in an ethical machine that combines consequentialist and deontological principles, knowing "which form the more absolute rules of deontological ethics" (Gerdes & Thornton, 2016, p. 88).

The problem with ethical issues in autonomous vehicles is well illustrated by the following example. The US National Highway Traffic Safety Administration (NHTSA) published two versions of a policy and guidance document to guide manufacturers in developing, testing, and deploying autonomous vehicles before they produce commercial versions to be used on public

[8] IEEE Code of Ethics accessed November 28, 2017 from https://www.ieee.org/about/corporate/governance/p7-8.html

roads (NHTSA, 2016; 2017). In the first version published in 2016, "ethical considerations" were in its scope; however, NHTSA removed such considerations from the second version published in 2017[9]. They did this for a variety of reasons. One reason the NHTSA offered is that ethical considerations in autonomous vehicles comprise an area that needs further discussion and research; therefore, it is too early to include ethical considerations in the policy and guidance document. Furthermore, according to NHTSA there is no general agreement on acceptable ethical decision-making processes in autonomous vehicles because they are not yet fully understood and there are no criteria to evaluate them. Therefore, the NHTSA wants to work with different sectors such as the automobile industry, governments, and regulators to conduct further research, create a framework to address ethical considerations, and enhance transparency in autonomous vehicle decision-making.

So, why is ethics in autonomous vehicles (AV) not yet understood? Here are some ethical questions: In an unavoidable crash, does an AV decide to minimize deaths or costs? Does an AV prioritize its passengers at all costs? Should AVs have different modes of ethical operation such as "self-sacrifice" or "passenger safety"? Should AVs passengers be counted? Should AVs consider the culture of the society in which they are operating? Can AVs detect and distinguish between children, seniors, and animals, and if so, how should they behave towards them? What type of ethical principles should be embedded in AVs? What are AVs' relative priorities, and are these priorities explicit? Can passengers change these priorities? Who is responsible for failures and decisions?

[9] NHTSA. (2017). Automated Driving Systems 2.0: A Vision for Safety. Retrieved November 24, 2017 from https://www.nhtsa.gov/manufacturers/automated-driving-systems

Some answers to these questions have been formulated as ethical rules for autonomous vehicles. The German Federal Ministry of Transport and Digital Infrastructure (BMVI) developed twenty ethical rules that serve as guidelines for autonomous vehicles (BMVI, 2017). The rules are divided into ten groups: one general rule, three rules on the ethical benefits of autonomous vehicles, five rules on unavoidable crash situations, two rules on responsibility, one rule on public information, two rules on safety and security, one rule on data protection, two rules on human-machine interfaces, two rules on machine learning, and one rule on public awareness (Luetge, 2017). The most important rules are those which relate to ethics in unavoidable crash situations. In rule 7 of these guidelines, autonomous vehicles must be programmed to first prioritize humans by preventing them from being injured or killed — thus implicitly making damages to animals and property acceptable. Rule 8 states that choosing between two human lives cannot be standardized or programmed. Therefore, in rule 9, autonomous vehicles may reduce the number of injuries but it is prohibited for autonomous vehicles to discriminate based on personal features such as age or gender. In addition, autonomous vehicles must not sacrifice innocent people who are not involved in the crash.

3.7.1. Autonomous Vehicles' Levels of Autonomy

As of the writing of this thesis, most technologically advanced vehicles that are used on roads are equipped with Advanced Driver Assistance Systems (ADAS) which help drivers to drive safely. ADAS include, for example, Adaptive Cruise Control, Emergency Braking, Pedestrian Detection, Collison Avoidance, Traffic Sign Recognition, Lane Departure Warning, Lane Keeping Assist, Cross Traffic Alert, Park Assist, Automatic Parking, Surround View, Blind Spot Detection, and Rear Collision Warning. In some respects, these capabilities are just more advanced versions of the "Eco mode" buttons that have existed in many vehicles for some

time. These Eco mode buttons change certain driving characteristics of the vehicle that lead to reduced fuel consumption[10]. The Eco mode has a positive impact on environment and society because it enhances the fuel efficiency of the vehicle, effectively reducing pollution[11]. Fuel economy in vehicles requires less gas to drive long distances with less emissions, which results in less pollution and less costs on gas. Therefore, even the simple decisions made by Eco mode systems have ethical impacts on individuals and society as whole. We have already begun to adopt vehicle technologies that help to reduce pollution, accidents, injuries, deaths, direct costs, and indirect costs. Even a feature as simple as Eco mode has positive ethical impacts on individuals, societies, and countries.

Vehicles which have all these types of technologies — including those in ADAS — could be called partially autonomous, which is level 2 based on the Society of Automotive Engineers (SAE International) Automation Levels (see Table 4). In addition to these technologies, some vehicles have an autopilot system which could be used on highways to keep the vehicle driving in its lane at a certain speed. This system uses some ADAS technologies, especially Adaptive Cruise Control and Lane Keeping Assist. Actually, vehicles that have the autopilot feature on highways could be called Conditionally Autonomous, which is the SAE's level 3 in Table 4. Despite these Conditionally Autonomous systems, the driver has complete control even while these systems are functioning; consequently, the driver is responsible for the vehicle's actions. The difference between ADAS and autonomous vehicles is who has responsibility (Johansson & Nilsson, 2016). Almost all of the vehicles that have autopilot systems require a driver to turn on the system and keep watching it in case the situation needs human intervention; for

[10] Boldt, D. (2016, November 15). What's This 'Eco' Button in my New Car? Retrieved December 6, 2017 from https://www.autoblog.com/2011/01/28/car-eco-button/

[11] UCS. (2010). *The Road Ahead. Union of Concerned Scientists*. Retrieved from https://www.ucsusa.org/assets/documents/clean_vehicles/the-road-ahead.pdf

example, the driver's hands must be on the steering wheel at all times during auto-piloting, otherwise the system will emit warnings until the driver puts his or her hands back on the steering wheel. If there is no response from the driver, then, in some vehicles, the autopilot system will gradually slow down the vehicle and bring it to a complete stop.

Some people say that fully autonomous vehicles will drive themselves, just like autopilots do in commercial airplanes. However, one problem that autonomous vehicles will face is operating in unstructured environments that could encounter unpredictable situations and changing weather (M. Wagner & Koopman, 2015), especially when many other objects share the same environment with the autonomous vehicles. Färber (2016) observes that "…road traffic is self-organizing… unlike air traffic" which is structured, guided, and controlled (p. 125). Färber adds that airplanes are controlled by strict rules; moreover, external supervision helps airplane pilots to operate their airplanes.

Table 4 SAE International Automation Levels (SAE, 2014)

SAE Level	Name	Narrative Definition	Execution of Steering and Acceleration/ Deceleration	Monitoring of Driving Environment	Fallback Performance of Dynamic Driving Task	System Capability (Driving Modes)
Human driver monitors the driving environment						
0	No Automation	-The full-time performance by the human driver of all aspects of the dynamic driving task, even when enhanced by warning or intervention systems	Human Driver	Human Driver	Human Driver	n/a
1	Driver Assistance	-The driving mode-specific execution by a driver assistance system of either steering or acceleration/deceleration using information about the driving environment with the expectation that the human driver performs all remaining aspects of the dynamic driving task	Human Driver and System	Human Driver	Human Driver	Some Driving Modes
2	Partial Automation	-The driving mode-specific execution by one or more driver assistance systems of both steering and acceleration / deceleration using information about the driving environment with the expectation that the human driver performs all	System	Human Driver	Human Driver	Some Driving Modes

		remaining aspects of the dynamic driving task				
Automated driving system ("system") monitors the driving environment						
3	Conditional Automation	-The driving mode-specific performance by an automated driving system of all aspects of the dynamic driving task with the expectation that the human driver will respond appropriately to a request to intervene	System	System	Human Driver	Some Driving Modes
4	High Automation	-The driving mode-specific performance by an automated driving system of all aspects of the dynamic driving task, even if a human driver does not respond appropriately to a request to intervene	System	System	System	Some Driving Modes
5	Full Automation	-The full-time performance by an automated driving system of all aspects of the dynamic driving task under all roadway and environmental conditions that can be managed by a human driver	System	System	System	All Driving Modes

To understand more about these levels, roles, and timing, NHTSA made simplifications to the SAE International Automation Levels (see Table 4) as shown in Table 5. As we see, ADAS are in levels 1 and 2. In levels 3 and 4 an Automated Driving System (ADS) takes partial control of the vehicle, and in some circumstances (e.g. autopilot) takes complete control over it. In level 5 ADS controls all of the vehicle's systems. There are no drivers in these types of fully autonomous vehicles; there are only passengers.

Table 5 Levels of Automation Explanation[12]

Level of Automation	Who Does What and When
Level 0	The human driver does all the driving.
Level 1	An advanced driver assistance system (ADAS) on the vehicle can sometimes assist the human driver with either steering or braking/accelerating, but not both simultaneously.
Level 2	An advanced driver assistance system (ADAS) on the vehicle can itself actually control both steering and braking/accelerating simultaneously under certain circumstances. The human driver must continue to pay full attention ("monitor the driving environment") at all times and perform the rest of the driving tasks.
Level 3	An Automated Driving System (ADS) on the vehicle can itself perform all aspects of the driving task under certain circumstances. In these circumstances, the human driver must be ready to take back control at any time when the ADS requests the human driver to do so. In all other circumstances, the human driver performs the driving tasks.

[12] NHTSA. (2017, September 9). Automated Vehicles for Safety. Retrieved November 25, 2017 from https://www.nhtsa.gov/technology-innovation/automated-vehicles-safety

Level 4	An Automated Driving System (ADS) on the vehicle can itself perform all driving tasks and monitor the driving environment — essentially, do all the driving — in certain circumstances. The human driver need not pay attention in these circumstances.
Level 5	An Automated Driving System (ADS) on the vehicle can do all the driving in all circumstances. The human occupants are passengers only and need never be involved in driving.

3.7.2. Autonomous Vehicle as Trolley Problem

Gerdes and Thornton (2016) maintain that Asimov's Laws are not sufficient for autonomous vehicles because they do not contain a complete ethical framework. As a solution, the authors introduce three Asimov-like laws that prioritize human lives plus one deontological rule: (1) An automated vehicle should not collide with a pedestrian or cyclist; (2) An automated vehicle should not collide with another vehicle, except in cases where avoiding such a collision would conflict with the first law; and (3) An automated vehicle should not collide with any other object in the environment, except in cases where avoiding such a collision would conflict with the first or second law. Finally, the deontological rule is: "An automated vehicle must obey traffic laws, except where obeying such laws would conflict with the first three laws" (Gerdes & Thornton, 2016, p. 97).

Since 2015 there has been considerable debate about how an autonomous vehicle would make decisions in trolley-problem like situations (Gogoll & Müller, 2017). Connecting the trolley problem to ethical decision-making by autonomous vehicles can be seen with increasing frequency in both news articles and in academic research because of the rapid growth in the AV industry (Millar, 2017) (Lin, 2016) (Goodall, 2016) (Wallach & Allen, 2009).

However, it is important to note that trolley problem was designed to reason about utilitarianism not to deal with a practical problem. Therefore, always it is used principally for

two reasons, 1) to demonstrate how the ethical decision-making process works and 2) to test people reactions toward using ethical judgment to make decisions that maximize utility.

Suppose an autonomous vehicle is in an unavoidable crash situation and the vehicle's decision-making system must calculate and evaluate the situation to make an ethically optimal decision. The decision-making process of an autonomous vehicle in a crash situation is a concrete example of utilitarianism in action. Several articles (Goodall, 2014; Johansson & Nilsson, 2016; Lin, 2014; Millar, 2014; Nyholm & Smids, 2016) discuss how autonomous vehicles would make decisions during crashes and how they are programmed to reduce damage and harm.

The prevalence of the trolley problem for machine ethics and autonomous cars in particular is illustrated well by the thought-experiments shown on MIT's Moral Machine website (http://moralmachine.mit.edu). This is an effective web platform for experimenting with different scenarios, exploring the moral dilemmas that could present themselves to autonomous vehicles in unavoidable crash situations, and evaluating the human moral judgments in these scenarios. In each scenario, participants need to choose between two harms based on their ethical preferences. Such scenarios on this platform are normally used to analyze the psychology of human preferences for the choices that a human would want an autonomous vehicle to make. However, those scenarios can also be a source of scenarios that a machine-ethics module would have to be able to reason about to satisfy these human preferences and to conform to social expectations.

A variant of the trolley problem, called the tunnel problem, was introduced by Jason Millar (2016):

> *Sarah is travelling along a single-lane mountain road in an autonomous car that is fast approaching a narrow tunnel. Just before entering the tunnel a child errantly runs into the road and trips in the center of the lane, effectively blocking the entrance to the tunnel. The car is unable to brake in time to avoid a crash. It has but two options: hit and kill the child; or swerve into the wall on either side of the tunnel, thus killing Sarah. It continues straight and sacrifices the child.* (p. 787)

This scenario was posted on a webpage in 2014 (Millar, 2014) with a poll (Robohub, 2014a) that asked the participants this question: "If you find yourself as the passenger of the tunnel problem described above, how should the car react?" The results show that 64% of respondents chose to kill the child rather than kill themselves (Robohub, 2014b). If the autonomous vehicle had full control of itself and faced the same dilemma, should it kill the child or kill its passenger? What if the autonomous vehicle chose to avoid the child and kill its passenger but the passenger has the ability to take control of the vehicle and kill the child instead? In these situations, would the vehicle, as an autonomous robot, violate Asimov's Laws and harm a human by obeying human orders? Or should the vehicle be fully autonomous and not have any human override capability that would enable the passenger to control it?

In the same poll, another question asked participants: "Who should determine how the car responds to the tunnel problem?" To answer this question, the participant had to choose from among four options: Passengers, Manufacturer/Designer, Lawmakers, and Other. Forty-four per cent of participants said that the passengers of the vehicle should decide if it should swerve or not, 33% of participants chose lawmakers, 12% chose manufacturers or designers, and 11% chose other. This raises another question: should autonomous vehicles decide what to do based on the passengers' ethical beliefs? Gogoll and Müller (2017) call this Personal Ethics Setting (PES), which allows individuals to decide what the ethical rule is that controls the vehicle. On the other hand, the Mandatory Ethics Setting (MES) is an ethical rule that has been agreed

upon by a society to be embedded in autonomous vehicles. In fact, most participants in the poll wanted PES which grants individuals the autonomy to choose how they want to live their lives. Yet PES as a right for individuals should not give individuals the right to end others' lives, which, from a purely utilitarian point of view, is not necessarily optimal. Gogoll and Müller (2017) discuss more drawbacks of PES showing how it may allow people to target minorities or certain groups over others as a type of algorithmic bias. In addition, PES would place too heavy of a responsibility on individuals by expecting them to make ethical choices that ought to be the responsibility of manufacturers and governments. Millar (2017) adds that because of the time it takes between detecting an object and crashing into it in an unavoidable crash situation, individuals will not have the time to make a decision. Therefore, Gogoll and Müller (2017) argue that MES is in the best interests of society.

Returning to the question of unavoidable accidents, Goodall (2014) introduced another scenario to illustrate another dilemma that is discussed further by Lin (2014) and referred to by Millar (2017) as the *helmet problem*. In this scenario an autonomous vehicle is in an unavoidable crash situation and must choose between two actions: turn left or turn right. On the left, there is one motorcyclist who is wearing a helmet; and on the right, there is another motorcyclist who is not wearing a helmet. What should the autonomous vehicle do? This ethical dilemma has different variations and autonomous vehicles are eventually going to face many of these variations in the future. In a utilitarian model for ethical behaviour, autonomous vehicles faced with an inevitable collision will choose the safest way to crash that minimizes harm and damage. However, Goodall (2014) added that this kind of crash calculation by the autonomous vehicle might lead to undesirable results. For example, if the autonomous vehicle faces an unavoidable crash with these two motorcyclists, then it will choose to crash into the

helmeted one instead of helmetless one because the autonomous vehicle minimizes damage and harm. However, this is unfair "not only because of discrimination but also because those who paid for safety were targeted while those who did not were spared" (Goodall, 2014, p. 62). Lin (2014) argues that crashing into the helmeted motorcyclist would essentially punish him/her for acting responsibly.

In the same context, autonomous vehicles in an unavoidable crash would make a decision that minimizes the harm by targeting the safest object to hit, such as an expensive vehicle which has more safety features to protect its passengers. If that happened, members of the public, especially those people who pay more to be safe, would learn not only to not trust autonomous vehicles but also to avoid them because they would be potentially more dangerous to them than to others. Detecting an object on the road would not necessarily mean that the autonomous vehicle would avoid it, because the vehicle would also be gathering data about other objects and comparing which is more valuable to the vehicle's system. Ultimately the autonomous vehicle could chose to run over the less valuable object or sacrifice its passengers (Surden & Williams, 2016).

The ability to dynamically compare potential targets in such crash situations has two disadvantages. On the one hand, people who live in neighbourhoods in which autonomous vehicles are driving will not care about the safety equipment in their own vehicles; at the same time, they will try to avoid being in the autonomous vehicles' field of vision. In such circumstances, it could be that autonomous vehicle behaviours would change the way that people behave and negatively affect their trust in such technologies. On the other hand, people will not buy autonomous vehicles if they have been shown to sacrifice their passengers in unavoidable crashes.

3.7.3. Autonomous Vehicle Communications

On the road the communication between pedestrians and drivers must be maintained. For example, if a pedestrian is on a crosswalk, he/she must be able to predict the driver's likely behaviour by observing some indicators such as car speed, rate of deceleration, and extent of the driver's awareness of his/her/its surroundings. Next, the pedestrian will communicate with the driver using eye contact to make sure that the driver is aware of the situation and prepared to allow the pedestrian to cross (Surden & Williams, 2016). In some cases, drivers wave to pedestrians as a signal for them to cross the street. Waving is a very clear indicator of the driver's awareness of the pedestrian's presence and intentions.

If there are pedestrians who want to cross the street in front of an autonomous vehicle, how will the pedestrians know that the AV will be stopping to allow the pedestrians to cross? In this case, there can be no eye contact because there are no drivers in AVs. Slowing down is not a sufficient indicator for pedestrians, because the AV may be slowing down before the crosswalk for another reason. Consequently, how is a pedestrian to know that the AV has detected them and is aware of their presence? (Surden & Williams, 2016). Thus, an AV needs to be able to demonstrate its awareness to others on the street who interact with it. Other vehicles (either AVs or human-driven vehicles), pedestrians, and cyclists must know that AVs are aware of them and detect them in their systems. However, even if the autonomous vehicle is able to show that it detects others, there is no guarantee that it will let them cross or avoid them in an unavoidable crash (Surden & Williams, 2016). Consequently, Surden and Williams (2016) introduce three suggestions to make autonomous vehicles work more effectively: (1) educate citizens about what AVs are capable of doing; (2) communicate with people who are detected; and (3) show them what the AV will do next.

To illustrate the first two suggestions, I draw from the following examples: In 2015, a group of people tried to demonstrate and test the Pedestrian Detection and Emergency Braking systems used by Volvo in its XC60 model. In the demonstration, two people stood in front of the car. First the driver backed up, stopped, and then drove quickly toward the two people who were standing in front of the car. These two people trusted the detection system to sense their presence and to brake by itself. However, the car did not stop at all and instead crashed into the people. In fact, according to Volvo's spokesman, the people who bought the car did not opt to purchase the Pedestrian Detection system which costs more than the vehicle they bought[13]. Therefore, it is important to educate society about vehicles' capabilities so that potential consumers will not overestimate them. According to Vallor and Bekey (2017), "A related ethical concern … [is that] humans greatly overestimate or rely unduly upon the capabilities of computerized systems". Raising people's awareness of the danger of system misuse is very important (IEEE, 2017).

The second example, which relates to the second suggestion, is that some semi-autonomous vehicles (level 2) contain a Blind Spot Detection system, which helps drivers by warning them about objects in their blind spot. In these vehicles, when sensors detect an object in the blind spot zone, a small indicator light (e.g. yellow or red triangle) located on the side mirror will illuminate to alert the driver (NHTSA, 2014). This light can also be seen by drivers of other vehicles, including the driver of the vehicle located in the blind spot, and therefore serves as a type of communication between the drivers: one vehicle detects the presence of the other and informs both drivers about its presence in a blind spot zone. Other types of vehicles have a

[13] Hill, K. (2015, May 26). Volvo says horrible 'self-parking car accident' happened because driver didn't have 'pedestrian detection'. Retrieved December 1, 2017 from https://splinternews.com/volvo-says-horrible-self-parking-car-accident-happened-1793847943

similar indicator light but it is located inside the side mirror (on the window frame) which can only be seen by the vehicle's driver. In this case other drivers in the blind spot zone do not know that their vehicles have been detected, and these drivers do not know if the driver of the other vehicle is aware of their presence or not. Accordingly, autonomous vehicles need such indicators to be visible to other human drivers on the road. Vehicle-to-vehicle technology could solve such problems: vehicles could communicate with one another about their situations including their speed, location, brake status, and other data to improve driving safety on roads, reduce traffic-congestion problems, and avoid accidents (Knight, 2015).

Färber (2016) explains that there are three types of road user communication: schema formation, anticipatory behaviour, and non-verbal communication. In schema formation, for example, the behaviours of an elderly man differ from that of a child, so a driver can tell from a distance whether a pedestrian is an old man or a child, or maybe physically disabled, blind, or Intoxicated. In anticipatory behaviour, for example, a pedestrian reaching a crosswalk is an indicator that the person intends to cross the street. Non-verbal communication is divided into two classifications of facial expressions (e.g. eye contact) and gestures and body movements (e.g. waving). Färber (2016) adds that all of these ways of communicating are understood in relation to their context. For example, a police officer waving to redirect traffic is understood by human drivers as legal instructions that must be obeyed. But if a random pedestrian imitates the gestures and body movements of a police officer, will autonomous vehicles follow these instructions? Assuming that autonomous vehicles are able to understand the intended communications behind these gestures, would they be able to distinguish between a police officer's gestures and a civilian's imitation of them? Furthermore, it could be that police officers' gestures and body movements for directing traffic are vary greatly from one country

to another given that there are no universal gesture rules. Consequently, autonomous vehicles would need to learn about the culture of the society in which they would be operating. Different environments have different communication methods which entails that autonomous vehicles companies must consider cultural variations. In that situation, we may need to develop vehicle-to-human technology that is similar to vehicle-to-vehicle technology. Vehicle-to-human technology would allow an autonomous vehicle to communicate with humans (drivers of other vehicles, pedestrians, and cyclists) by telling them about its current state and its next intended move.

3.7.4. Societal Issues

There are other societal issues inherent to the behaviour of robotic systems such as the bullying of autonomous vehicles. Some people foresee that autonomous vehicles might be bullied (LSE, 2016). This is actually a real issue that could cause a lot of problems. Aggressive drivers could cut off autonomous vehicles or stop in front of them to hamper or distract them. Dietmar Exler, CEO of Mercedes-Benz USA, worries that humans will bully autonomous vehicles[14], and Eruj Coelingh, Volvo's senior technical leader, shares that fear. Thus, the first generation of autonomous vehicles in UK will be unmarked because if they were identifiable, people would test them by cutting them off, standing in front of them, or performing harsh behaviours toward them[15]. In fact, this has happened already with Google's autonomous vehicle. In early testing it was unable to get through a four-way stop because it was waiting for other drivers to come

[14] Mitchell, R. (2016, November 15). Human drivers will bully robot cars, says CEO of Mercedes-Benz USA. Retrieved December 3, 2017 from http://www.latimes.com/business/la-fi-hy-live-updates-2016-la-auto-show-human-drivers-will-bully-robot-cars-1479247249-htmlstory.html

[15] Connor, S. (2016, October 30). First self-driving cars will be unmarked so that other drivers don't try to bully them. Retrieved December 3, 2017 from http://www.theguardian.com/technology/2016/oct/30/volvo-self-driving-car-autonomous

to a complete halt before moving forward. Human drivers discovered that if they kept moving forward, they could prevent the Google AV from making it through the intersection[16].

Brooks (2017) believes that the concern with AVs from the unexpected consequences that arise during the development and testing phase will delay the commercial deployment for a long time. He illustrates this with some examples of behaviours by autonomous vehicle owners which highlight some of the social problems involved with introducing AVs on the roads. In the first example, owners or passengers of autonomous vehicles may take advantage of their autonomous control and get out of their vehicles before they are properly parked. They would know that their vehicles will be able to take care of whatever situation they are in and move themselves out of, for example, an illegal parking position, if necessary. While this may work well, Brooks (2017) adds that this type of behaviour may have an effect on other people and slow them down. In the second example, Brooks (2017) considers a family with two autonomous vehicles who wants to attend an event in a location where there are very few parking spots. The family members may take advantage of their circumstances by sending the first AV earlier in the day to find and park in a spot that is close to the event. Later, they would use the second AV to drop them off at the event and send it home directly. Once the event is over, they would use the first vehicle that is waiting for them at the parking spot to pick them up. The problem of autonomous vehicles occupying parking spots for a long time could increase costs and inconvenience other drivers. This type of technology could therefore affect

[16] Richtel, M., & Dougherty, C. (2015, September 1). Google's Driverless Cars Run into Problem: Cars with Drivers. Retrieved November 30, 2017 from https://www.nytimes.com/2015/09/02/technology/personaltech/google-says-its-not-the-driverless-cars-fault-its-other-drivers.html

the lives of citizens in unintended ways in the same way that the Google maps feature of offering alternate routes to avoid congested roads has affected residential neighbourhoods[17].

In the same context, since autonomous vehicles can detect and respond to road signs and street markings, people could vandalize or repaint road signs to bully AVs. For example, a man in China painted new arrows to redirect the traffic in certain places to make the bus he rides every day drive him to work faster[18]. In another example, a group of people from a Toronto neighbourhood redesigned a poorly designed intersection by using chalk and leaves[19]. Whether motivated for bad or good reasons, how should autonomous vehicles respond to these changes?

3.8. Machine Learning

Machine learning is a programming method to enable a machine to learn to improve its performance by using examples or past experiences (Alpaydin, 2010; N. J. Nilsson, 1996). To enable an autonomous robot to make ethical decisions as a result of experiences would require a machine learning method. There are three types of machine learning: supervised learning, unsupervised learning, and reinforcement learning (Alpaydin, 2010). In *supervised learning* the robot's supervisor teaches the machine to correlate inputs and outputs by using labelled examples so that it may perform these labelling with new inputs and outputs. In *unsupervised learning* the robot obtains only the input and, without a supervisor, learns what input patterns are associated with given outputs. In *reinforcement learning* the robot learns that a certain goal in a dynamic environment can be achieved by performing a sequence of actions. IBM Watson,

[17] https://productforums.google.com/forum/#!topic/maps/L6Zrim9jbXM

[18] Shen, A. (2017, November 30). Chinese Man Repaints Road Markings to make his Commute Quicker. Retrieved December 3, 2017 from http://www.scmp.com/news/china/society/article/2122252/chinese-man-repaints-road-markings-make-his-commute-quicker

[19] CBC News. (2017, December 1). Using just chalk and leaves, Toronto residents re-imagine 'poorly designed,' 'dangerous' intersection. Retrieved December 3, 2017 from http://www.cbc.ca/news/canada/toronto/dave-meslin-sidewalk-toronto-chalk-leaves-1.4427663

for example, used several of these machine learning techniques to play Jeopardy (Ferrucci et al., 2010). It was trained with many questions and known answers (supervised learning) and also used reinforcement learning to learn game strategies. In another example, AlphaGo developers used supervised learning to train the system how to play Go, while AlphaGo Zero used unsupervised learning to play with itself[20]. Autonomous robots are also good candidates for being trained by reinforcement learning applications because they operate in dynamic environments which require a chain of actions that must be performed in sequence in order to reach a goal.

Why do we use machines that learn? Nilsson (1996) offers two reasons: (i) machine learning is often the most appropriate way to define some tasks, even if input and output are specified but the relationship between them is unknown; and (ii) some relationships between data are hidden because of its volume, but machine learning is able to recognize these relationships through data mining. Designers may not know how to build machines that perform tasks in a given environment but enabling a machine to learn enables the designers to know and discover how to perform these tasks. In addition, some tasks require a large amount of knowledge, which is impossible for humans to assimilate; however, machine learning is able to recognize that knowledge. Machine learning is an effective way for a machine to adapt to changing environments without programmers having to change its programming.

Because of the machine learning process, in some situations even the programmers or the machines' creators cannot understand how or why a machine has performed a particular task or made a specific decision (Surden & Williams, 2016). Vallor and Bekey (2017) concur: self-

[20] Hassabis, D., & Silver, D. (2017, October 18). AlphaGo Zero: Learning from scratch. Retrieved February 8, 2018 from https://deepmind.com/blog/alphago-zero-learning-scratch/

learning machines can behave in unpredictable ways, and it is possible that these behaviours cannot be understood or explained, even by their creators. This lack of understanding by machine designers poses a real ethical risk. Surden and Williams (2016) call some systems that learn "technologically opaque", because normal users cannot understand why such a system takes a particular action and cannot predict its future decisions and actions. However, a recent study conducted by UC Berkeley researchers introduces a method for robots to predict and imagine the future. These abilities allow robots to perform actions that lead to wanted results (Ebert, Finn, Lee, & Levine, 2017). Therefore, robots would have the ability to see their actions' consequences before these actions are performed. If this prediction ability is enabled for robot users, it could be a way to reduce the technology opacity. Robot users would have a clear visual understanding of the consequences of a robot's action before this action takes place. However, at present this method is limited to only allowing robots to predict the future only for a few seconds. Nevertheless, this is enough time for robots to learn by themselves without human intervention[21].

3.8.1. Algorithmic Bias

Discussing algorithmic bias here is essential because decision-making algorithms are used in a variety of domains from simple decision-making systems to complex profiling algorithms (Mittelstadt, Allo, Taddeo, Wachter, & Floridi, 2016). For example, decision-making algorithms are used in recommender systems to filter job applicants, to offer mortgages, and to predict crime rates. Mittelstadt et al. (2016) introduce six types of ethical concerns raised by such algorithms: *inconclusive evidence* leading to unjustified actions when the machine

[21] Israel, B. (2017, December 4). New robots can see into their future. Retrieved December 8, 2017 from http://news.berkeley.edu/2017/12/04/robots-see-into-their-future/

learning method produces uncertain knowledge; *inscrutable evidence* leading to opacity when transparency is needed because algorithms are unpredictable and difficult to control; *misguided evidence* leading to bias when the functionality of algorithms reflects programmers' values and results in biased decisions; *unfair outcomes* leading to discrimination when an algorithm's action affects certain people; *transformative effects* leading to challenges for autonomy; *transformative effects* leading to challenges for informational privacy; and *traceability* leading to moral responsibility when the machine fails and the accountability must be allocated.

There are several different sources of algorithm bias. For example, Hammond (2016) has identified five types of bias in artificial intelligence: data-driven bias, bias through interaction, emergent bias, similarity bias, and conflicting goals bias[22]. Similar types of bias were classified by Danks and London (2017): training data bias, algorithm focus bias, algorithm processing bias, transfer context bias, and interpretation bias. Some scholars discussed discrimination by algorithm and identified four types: input data bias, algorithm bias by itself, different context using bias, and feedback bias (Lepri, Oliver, Letouzé, Pentland, & Vinck, 2017). In general, algorithm bias comes from two sources: algorithm construction and the use of algorithms. Each source has different types of bias as mentioned earlier. For example, algorithm construction contains training data and training process biases, and algorithm usage bias includes interaction bias and context bias. Some of these forms of bias illustrate the need to include ethics modules in robots to protect individuals and society from autonomous machine-decisions that, for example, affect groups of people who are defined by their ethnic, cultural,

[22] Hammond, K. (2016, December 10). 5 unexpected sources of bias in artificial intelligence. Retrieved September 17, 2017 from https://techcrunch.com/2016/12/10/5-unexpected-sources-of-bias-in-artificial-intelligence/

and religious identity. Therefore, I focus on only four types of bias: those that are pertinent to the problem of transparency, explanation, and trust.

3.8.1.1. Training Data Bias

In this type of bias, a robot makes its decisions based on what it has learned from the data in the training process. For example, a system that evaluates the likelihood of future recidivism rates for parolees incorrectly predicts that black people are more likely than white people to commit subsequent crimes (Angwin et al., 2016). The problem with that system is that it relies on training data that is racially biased, such as residential stability (number of moves in the past twelve months), residential location, and incidence of poverty. Since the input data is not free of racial bias, it is not surprising that the output is not bias-free either. Missing or ignoring data during a robot's training process could also be part of the problem. If the robot does not recognize an object, it will be biased against it.

Another example of potentially unethical behaviour caused by data-driven bias is beauty.ai, an on-line beauty contest that used deep learning algorithms to judge contestants (Levin, 2016; Pearson, 2016). From among six thousand selfies submitted by contestants, the algorithm selected forty-four and grouped these into five age categories, judged to be pictures of the "most attractive" people. Of the winners, six were Asian, one had dark skin, and thirty-seven of the winners were white. Apparently, the machine had developed human-like biases through its learning process. The data-bias in this case depended not only on the training process (human judgments of "beauty") but also on the distribution of the training sample that was used in the database.

3.8.1.2. Interaction Bias

In bias through interaction, the robot learns to behave as a result of its interaction with people rather than from training data. The incident with Microsoft Twitter's chatbot Tay, for instance, illustrates the serious ethical consequences that result from the absence of modules for providing ethical guidance in decision-making algorithms. The Natural Language Processing AI and machine learning algorithms in Tay enabled it to adapt its verbal behaviour according to feedback from its conversations with users: the more Tay chatted, the more it learned what to say and how to respond during its interactions with people. However, less than twenty-four hours after its deployment, its followers on Twitter had taught it to tweet in ways that were rude, discriminatory, and racist (Bass, 2016). Tay's behaviour was unpredictable because no allowances were made in its design for a potentially "hostile" environment such as malicious humans' intent on teaching it to verbally misbehave. The danger to which the Tay bot alerts us is that intelligent learning systems that are designed without built-in ethical constraints, such as principles that govern ethical verbal behaviour, have the potential of causing great harm.

3.8.1.3. Context Bias

The transfer and deployment of algorithms in a context for which they were not originally designed may also produce biases. Using a robot that was built for a certain environment in another environment could lead to context bias. For example, using an autonomous vehicle in the United Kingdom that was originally designed to be trained and used in the United States would generate context bias because people in the UK drive on the opposite side of the road than people in the US (Danks & London, 2017). The same thing would happen to a social robot if it were used in a different culture. Social robots should have ethics that respect culture,

religion, traditions, and domestic rules. Such robots might be designed to make decisions and perform actions based on algorithms that perceive a specific culture, and if they moved to a different environment, they could violate local laws or offend users in different ways.

3.8.1.4. Emergent Bias

Emergent bias arises when robots use social network data or collective social data that manifests cultural values and societal knowledge. Hammond (2016) illustrates this bias by pointing to the way that Facebook's news algorithm recommends news that perpetuates belief "bubbles", thus creating "algorithmic version of confirmation bias".

Recommender systems in general manifest emergent bias. Several examples show the need for ethical guidance for the purpose of moderating these systems. A striking example is the recent discovery that Amazon's recommender could recommend to users who had added a harmless substance in their shopping basket the remaining ingredients that together make a bomb (Kennedy, 2017). The recommender behaved correctly and as expected, since the data that yielded this recommendation was certainly that of human beings who had purchased these items together. Yet these recommendations may also have been reinforced by the emergent bias of the "rich-get-richer" effect (Fleder & Hosanagar, 2009). When applied to recommending users to other users in social networks, recommender algorithms manifest the same effect (Su, Sharma, & Goel, 2016).

Without an ethical filter, such recommender systems that manifest an unforeseen emergent bias may make decisions that could affect human well-being, encourage people to break the law, or even cause harm to the environment. Some decisions they make need to be governed

by people by including a 'human in the loop', or by giving ethical guidance that leads these systems to make decisions that are modulated by ethical considerations.

3.9. Corporate Social Responsibility

Since the concept of Corporate Social Responsibility (CSR) plays an important role in the remainder of this thesis, I need to introduce it here in some detail. CSR can be approached from two different dimensions: ethical and economic. The ethical dimension refers to the corporate role in society toward stakeholders (such as government, employees, consumers, suppliers, and community). The economic dimension refers to the corporate financial role toward shareholders (such as corporate owners) (Renouard, 2010). Social responsibility is defined as "a business's obligation to maximize its positive impact and minimize its negative impact on society" (O. C. Ferrell, Hirt, & Ferrell, 2016, p. 36).

Before I discuss CSR, I must briefly discuss business ethics as it relates to CSR. Business ethics, which comprises an area of applied ethics, serves here to support my discussion of robotics companies' responsibilities toward societies. Essentially, applied ethics make up "the branch of ethics which investigates the application of ethical theories in actual life" (Tzafestas, 2016b, p. 15). As mentioned earlier, applied ethics contain different areas such as business ethics, robot ethics, legal ethics, media ethics, manufacturing ethics, automation ethics, and environmental ethics. Thus, business ethics forms an area of applied ethics that aims to guide decisions and actions of businesses (Collins, 2012; Marcoux, 2008). Other researchers define business ethics as a discipline that includes values, moral principles, and standards which guide business behaviours (J. Fraedrich, Ferrell, & Ferrell, 2011). For instance, accountability, integrity, and trust are examples of values. In addition, freedom, justice, and human rights are examples of moral principles. In the academic field, business ethics is defined as the study of

business morality which contains business values, market expectations, and practices (De George, 2012). Consequently, while business ethics is described from different perspectives, it generally focuses on the moral dimensions within corporations. These dimensions involve various people and organizations including shareholders, employees, customers, suppliers, competitors, government, community, and even environment. Business ethics focuses what is right and what is wrong in business decisions (Crane & Matten, 2007).

However, differentiating between right and wrong decisions in business is not an easy task, as conflict may at times exist between interests. For example, a moral decision could lead to a company spending more money than an immoral decision would. Occasionally moral decisions cost a lot of money, time, and effort. However, making money must not be the only immediate objective of any business: there are other important aspects of a business that must be addressed such as safety of customers and employees, quality of products, environment, and resources.

To apply ethics in business, a corporation should establish certain values to guide its decisions (Brusseau, 2011). These values include values which are inherent in declarations of human rights. In addition, a corporation must have a good understanding of its market, society, people, and regulations so as to be prepared to choose the most suitable action for each situation. Other authors state that two aspects should be addressed to apply ethics in business: profits and desires (J. Fraedrich et al., 2011). In their view, a business needs to make profits to survive in the marketplace; however, this desire for profits should not conflict with societal rights and needs. Thus, a business must find a way to balance its needs and society's needs.

According to Crane and Matten (2007), business ethics are important because: 1) the impact of businesses in society is greater than before and businesses need to take care of ethical aspects of their behaviours so as to affect society in positive ways; 2) businesses play several

major societal roles such as selling products and services, paying taxes, and creating employment; therefore, businesses can invest this involvement in beneficial ways; 3) principles of business ethics teach those working in businesses who may not be knowledgeable about morality how to make ethical decisions by providing the relevant knowledge to identify, diagnose, analyze, and solve problems; and 4) business ethics help those working in businesses to understand society from a professional standpoint by understanding human, societal, and environmental needs. Patrignani and Whitehouse (2015) suggest that technology and electronics companies have an especially big role in societies because their products are used extensively. Due to the internet and the ubiquity of communications technologies, electronics companies affect not only local society but also global societies. Thus, CSR is a crucial success factor which affects the performance of large technology companies.

According to O.C. Ferrell et al. (2016), although some people use the terms CSR and business ethics interchangeably, they are not the same thing. Business ethics focuses on work group decisions and behaviours, while CSR focuses on the impacts of business on society. In fact, there is more than one definition of CSR depending on one's perspective. According to Collins (2012), CSR is a method or strategy that is taken by a business to deal with decisions (legal, social, moral, or economic) and actions that affect different people and elements both within the business and external to the business such as stakeholders, society, and environment. Dahlsrud (2008) identified and analyzed thirty-seven other definitions in terms of five specific dimensions, namely the environment, society, economics, stakeholders, and voluntarism, and calculated the frequency of occurrence of these definitions on Google. Some definitions include all five dimensions; however, others cover fewer than five. Some definitions are attributed to corporations, and others are attributed to individual authors. For example, the Commission of the European Communities, as cited in Dahlsrud (2008), defines CSR as "A

concept whereby companies integrate social and environmental concerns in their business operations and in their interaction with their stakeholders on a voluntary basis" (p. 7). Based on Dahlsrud's classification, this definition is optimal because it includes all five dimensions of economics, stakeholders, voluntarism, the environment, and society. As well, this definition is the most frequently cited according to search hits on Google.

Although we can therefore conclude that companies can be evaluated in terms of social responsibility based on the five factors analyzed by Dahlsrud (2008), these five factors could alternatively be re-categorized along the three dimensions of people, the planet, and profit. The first dimension of people includes business employees and members of the surrounding community such as society and competitors. The planetary dimension consists of the environment, such as natural resources and sustainability. The third dimension of profit entails companies making money. These three factors or dimensions are called the Triple Bottom Line approach which was originally introduced by (Elkington, 1998). Others have used the acronym TBL (Crane & Matten, 2007) or 3Ps (Collins, 2012).

To conclude, the core concepts of robot ethics presented in this chapter include theories of ethics, responsibility, autonomy, machine learning and CSR. This survey provides a conceptual foundation with which to understand the overall picture of robot ethics and the relationships between its components. Additionally, one aim of this discussion has been to identify gaps in the literature that need to be filled, such as trust and transparency between robot corporations and society. These components form the basis of my framework, as presented in Chapter 4.

4. The Framework

The literature review (Chapter 3) and public opinions analysis (Chapter 5) indicate that *if* we are going to release autonomous machines that make decisions that have ethical consequences into the market, then we should equip these machines with the capability to make these decisions under the guidance of ethical principles. As well, robot manufacturers need to be guided by government regulations to determine how such autonomous robots should be designed to best serve the interests of members of society. In particular, governmental regulations need to instruct AI corporations about how to create ethical robots that enhance people's lives and gain people's trust.

In this chapter I present a trust framework for ethical robots, an ethical decision-making model, and a transparency model to answer the research questions[23]. The public opinion analysis in Chapter 5 provides evidence to support key elements of the transparency model described in this chapter. In this chapter I show how this model provides a more complete answer to questions 2 and 4. These questions are: (Q2) If humans are to trust robots to make ethical decisions, what are the conditions under which this trust is justified? (Q4) How do ethical decision-making algorithms need to be constrained in order to make them acceptable to the public? The transparency model presented in this chapter completes the answers to those two questions. However, before proposing the transparency model I need to build a general framework that shows the relationships among different components, including the component of transparency. This framework answers the third research question (Q3): If decisions by robots are made using ethical principles, what is the role of explanations in establishing trust

[23] Parts of this chapter have been published before in (Alaieri & Vellino, 2016; 2017)

and allocating responsibility? Thus, I call this a trust framework. Furthermore, to show the importance of ethics in the decision-making process of robots and to answer the first research question (Q1) — In what situations and why might it be inadvisable to enable robots to make ethical decisions? — I propose a decision-making model for ethical robots.

4.1. A Trust Framework for Ethical Robots

The framework aims to explain the relationships among different components of the trust network, including regulators, corporations, robots, society, computing elements, embedded ethical principles, and corporate social responsibility (CSR). In the framework described in Figure 7 we see five main groupings: a stakeholder group which includes regulators, corporations, and society as a whole; a computing group which contains algorithms and machine learning; an ethics group which includes ethical principles and the ethical decision-making process; corporate social responsibility which includes the elements of transparency and explanation; and sustainability with trust and reputation as the principal components.

To build a trustworthy robot that makes ethical decisions, *designers* can either use deterministic *algorithms* that embed *ethical principles* explicitly or use *machine learning* to teach the robot to perform *ethical decisions*. One advantage of using explicit ethical principles that are implemented algorithmically is that this easily satisfies the social need, including the need of end-users, to have ethical transparency. If robot builders know what types of ethical principles are being used in that robot, then they can publish these principles and explain how the robot will behave in any given situation. In particular, it satisfies the failure case — the need to know how and why the robot made a specific decision when an error occurs. Fulfilling these two requirements of transparency and explanation would encourage society, including

users, to trust such robots and trust the methods used by their manufacturers in robot design. The use of machine learning to teach a robot to perform ethical decisions presents a greater challenge for both transparency and for providing explanations, and hence for establishing the trustworthiness of robot behaviour. While machine learning methods produce remarkable results, they cannot yet provide "interpretations" that are sufficient to explain their behaviours. In addition to being a desirable attribute for end-users and the public, the trustworthiness of robot machines built by a manufacturer has an impact on the manufacturer's reputation. Trustworthiness is an outcome of the transparency of the decision-making process and the possibility of providing explanations for these decisions after a machine-generated decision has been made. Above all, regulators need to guide manufacturers to protect society from robots' potential failures and to encourage manufacturers to create more dependable robots.

It is important to emphasise that my primary focus in this thesis is on the ethics and CSR groupings. Therefore, in the next subsections I briefly explain all other groupings besides ethics and CSR and explain their impacts on the trust relationships between robots, users, and the public. However, for ethics and CSR groupings I further propose more detailed models which describe their internal processes. In the ethics grouping I present the ethical decision-making model which decomposes the stages of decision-making into their elementary components with a view to enable stakeholders to allocate the responsibility for robots' choices. In the CSR grouping I present the transparency model, which displays the relationship between disclosure, transparency, explanation, and social trust.

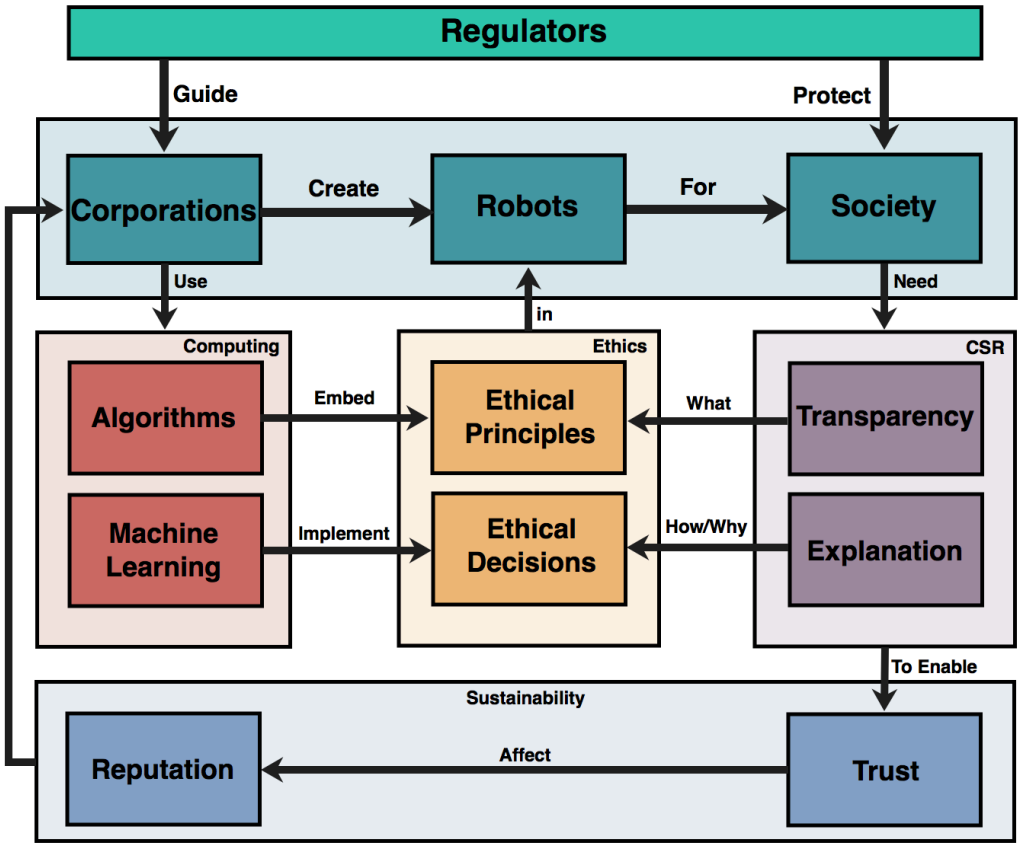

Figure 7 Trust Framework

4.1.1. Stakeholders

Governments, policy makers, corporations, users, and communities are the main stakeholders when it comes to ethically enabled decision-making machines. Simply put, government agencies, professional associations, and policy makers provide corporations with policies, regulations, recommendations, and laws that enable them to build safe robots and to maximize the robots' benefits to end-users and the public. Corporations, including designers, engineers, programmers, and developers, have responsibilities to follow this guidance in the development of robots and their deployment. Sometimes corporations have to regulate robots to be used ethically in their platforms. For example, social bots that work in social networks such as Twitter and Facebook may behave unethically, especially when they have been used by

corporations to post, share, and re-post the same or substantially similar content from different accounts simultaneously. Therefore, social networks have an ethical responsibility to control and guide such bots to prevent harm to users, including the dissemination of "fake news", be it information that influences an election or has an effect on public health (e.g. anti-vaccination mis-information). Owners and users of robots as well as people who live in communities with robots are all affected by corporations' products. The regulations exist to protect the public from robots' failures. Robots' failures could be divided into two categories: 1) failures that happen due to the absence of ethics (e.g. Microsoft Twitter's chatbot Tay) or algorithmic bias [see 3.8.1]; and 2) failures that happen because of an error in a robot's system which have nothing to do with ethics. However, in either case society needs there to be transparency in the operation of the machines and explanations for their behaviours.

4.1.2. Computing

The computing block in the framework refers to the software techniques that are used by the robots' decision-making processes. One way to establish trust is to ensure that given the same input to a machine's decision-making process, it produces same output by being deterministic. This is especially true if robots have ethical decision-making capabilities. Robots that have embedded deontological rules — e.g. to obey laws, to cause no harm, to not kill, etc. — need to obey these rules every time they make a decision or perform an action, particularly if that action has consequences that could cause harm. If their decisions are deterministic and predictable, then users will know what to expect. Whatever it is that counts as an "explanation" for the decisions that a machine makes, a machine that makes non-deterministic decisions — i.e. decisions that may produce different outcomes even if the initial conditions are the same — would need to provide a lot more detail about how it made its decisions in comparison to

what is required from a deterministic robot. This is a significant problem with machine-learning robots that learn from experience.

On the other hand, requiring machines to make only deterministic decisions — i.e. not be able to learn from experience — even if those decisions are ultimately unethical may preclude these machines from improving ethically. For example, an autonomous vehicle which only has deontological rules that require it to obey traffic laws, and hence never cross through red lights, may not be able to override the decision to stop in a situation when it should violate traffic laws for the greater good (e.g. if driving through a red light is necessary to make way for an ambulance). The outcome of not making way for the ambulance may result in more harm (possibly even death) than the outcome of violating a deontological rule that was based on a utilitarian calculation that resulted from a machine learning module designed to improve the decision-making process based on experience. It is therefore tempting to believe that it is better to have a learning system that improves with experience than it is to have a machine that makes deterministic decisions. Yet it is precisely the unpredictability of robots' decisions that makes it imperative to impose known ethical constraints that govern robots' actions. There must be some degree of certainty that machines will not contravene people's values.

Indeed, in this thesis I do not try to design algorithms and machine learning solutions to solve the problem of embedding ethical principles in a robot. While this is not in the scope of my thesis, this could be future work for others to improve the framework. Here, I am interested in understanding the role of algorithms and machine learning as important factors in cementing the trust of users and citizens in such decision-making machines.

4.1.3. Ethics

For ethics-enabled robots to be trustworthy, their actions must be explainable. Therefore, the process by which the decisions led to these actions must be transparent. Actions that are the result of a decision-making process in which consequences are evaluated using a consequentialist method or which are assessed by deontological principles are all in need of transparency. If ethical decision-making processes are opaque, i.e. performed by a "black box" whose inner workings are inaccessible or hard to interpret, then end-users will rightly be concerned about who can be held accountable when mistakes are made. This is particularly true of self-learning robots whose ethical behaviours are learned. It is worth mentioning that some self-learning decision-making robots (e.g. AlphaGo Zero) are not yet explainable in human terms (Silver et al., 2017).

Existing intelligent decision-making machines, whether driven by deterministic algorithms or machine learning methods, may not always as effective at making good decisions as they could be. The algorithms that are relied on by machines to perform actions are developed by humans and may contain errors or bugs. Furthermore, supervised machine learning systems that are trained by humans may manifest training biases. The outcomes may either be entirely incorrect, be correct but only under some conditions, or have unanticipated ethical impacts on some groups of people. Such intelligent decision-making algorithms have already become part of many people's daily lives. For example, these decision-making algorithms are used in search engines, social media suggestion and recommendation systems, decision-making systems, and prediction tools. Until recently these decision-making systems had not resulted in obviously important ethical consequences such as harming someone on the road or

influencing the outcomes of elections. Therefore, there has been little scrutiny of how these decision-making systems or their developers should be held accountable for their decisions.

Following the Sense-Think-Act robotics paradigm in (Beer et al., 2014; Parasuraman et al., 2000; Vanderelst & Winfield, 2016), I distinguish ten stages in the decision-making process of a robotic device and identify four of those stages as critical to performing autonomous ethical decisions as illustrated in Figure 8. The ten stages in this conceptual model are themselves divided into two main clusters: the Planning Analysis cluster and the Ethical Guidance cluster. The main function of planning analysis is to determine whether an action-choice requires an individually situated ethical analysis or whether it can short-cut to a decision. If it does require ethical analysis, then the four stages in the ethical guidance cluster come into play.

Each stage is best illustrated using example robots. Consider a delivery robot such as "6D57" by Star Ship technologies (Lonsdorf, 2017) that delivers packages from point A to point B. Point A could be a pizza store and point B could be a customer's address. The delivery task is composed of an ordered set of subtasks which must be performed in sequence to reach the overall goal — getting the package, navigating on sidewalks or streets to the destination, notifying the customer, delivering the package, and returning home. Another example I use below is that of a chess-playing program that selects a winning move from among alternatives. Note that in situations where there are no specific ethical decisions that a robot needs to make, it may nevertheless need to choose from among alternative courses of action. Such non-ethical decision-making obeys a similar process to the ethical guidance process described below except that the "objective function" — the quality that the decision aims at maximizing — is different. Furthermore, some of the stages in an ethical guidance cluster are only proposed and

may not have a real-world example, so I continue to use the same delivery robot example to illustrate the meaning.

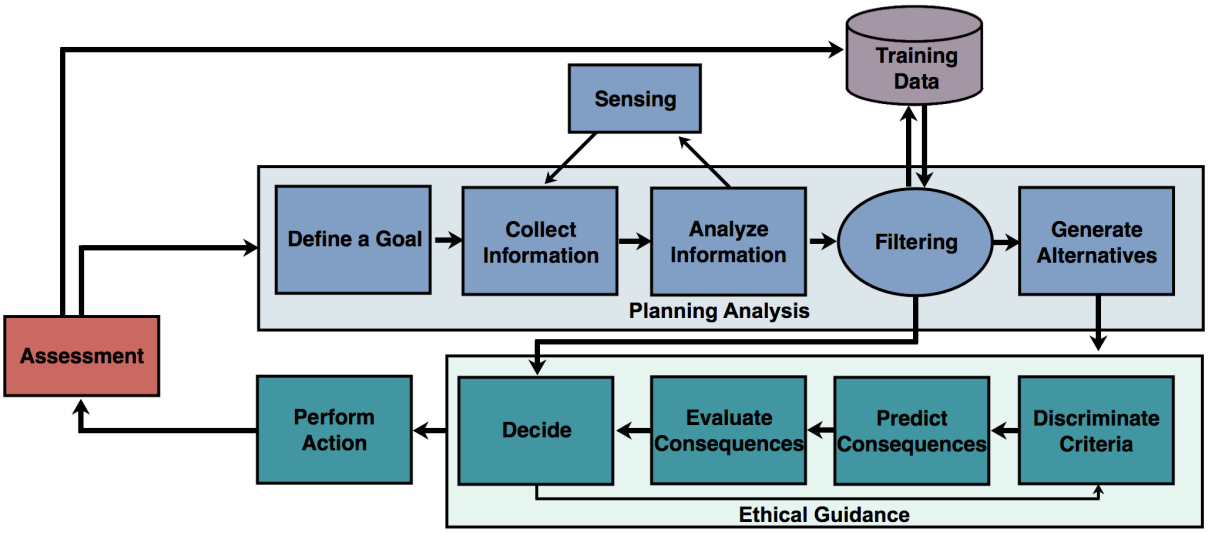

Figure 8 Decision-Making Model for Ethical Robots

A. Planning Analysis

1. Defining a Goal.

In the "6D57" delivery robot example, the robot's main goal is to deliver the pizza to the customer and return to the store. In this type of robot, the goal is built-in: delivery robots have no other function. In contrast, a multi-functional robot could have programmable goals and a fully autonomous robot could potentially define its own goals.

2. Collecting Information.

Robotic systems typically acquire information from sensors, networks, telemetry, and humans. The information they need is obtained from sound, light, physical contact, and other sources of information input methods. The delivery robot would obtain the customer's address from a

data source, determine its own location from a GPS, use its sensors to recognize obstacles and identify possible paths in the environment, and use a map to navigate them.

3. Information Analysis.

Robotic systems then analyze this input information to create real-world representations of the objects they need to reason about, synthesize and integrate this data to make predictions, and execute planning tasks. These two stages of collecting and analyzing information are the means of "sensing" the environment. These stages may need to be repeated several times to obtain the desired data. In the delivery robot, this may mean the identification of obstacles, which leads to the dynamic reconfiguration of the trajectory to its destination.

4. Filtering.

Once the information has been collected and analyzed, the robot system will then "filter" it. This filtering process occurs in machines that have an experiential learning component, even a rudimentary one. For example, if information collection and analysis has taken place once before and a decision has been made on that basis, then there may be no need to re-compute the decision-making process when similar or identical conditions arise again. In this case, it may be simpler to retrieve the prior decision and simply perform the appropriate action. For example, the delivery robot could recognize the entered address if this address is for a returning customer. In this situation, the delivery robot could navigate the prior path that was used the last time it delivered a package there.

5. Alternatives Generation.

If the filtering stage indicates that it is necessary to go through the process of making a decision, the robot then needs to produce a list of alternative actions from which to select and finally execute the selection process, i.e. to make a decision. The generation of feasible board positions in computer game play is a typical example of this stage. For instance, when a computer plays chess it generates dozens of alternative moves and considers thousands of alternative positions that may follow from selecting any one of these moves by using a process similar to the one outlined below for making ethical decisions. In the case of the delivery robot the "alternatives generation" stage would occur if there were alternative routes for reaching its destination; it may, for example, need to select which roads to cross, choose one from among several alternative parking spots, or decide at which time to notify the customer.

B. Ethical Guidance

The ethical guidance stages presented in Figure 8 govern the robotic systems' ethical decision-making. These stages embody ethical deontological and/or utilitarian principles, including the relevant laws, regulations, cultural values, and codes of conduct. Each stage is outlined below.

6. Discriminate Criteria.

Given a set of alternative actions from which to choose, the robotic system should now evaluate the generated alternatives according to deontological rules. Some alternatives may need to be eliminated either because they violate ethical codes of conduct or because they are illegal. For instance, in the case of the delivery robot, the need to obey the speed limits (both

the upper limits and the lower limits) may preclude certain routes. Similarly, parking spots reserved for pregnant women must be excluded from possible alternatives.

7. Consequences Prediction.

If the action to be selected needs to be evaluated according to consequentialist principles (stage 8), then the implications of each alternative will need to be computed. In the case of the chess-playing bot, this stage amounts to exploring the search space of possible subsequent board positions that may follow from each alternative, noting that the chess-playing bot is not making ethical decisions but is an example of consequences prediction. In the case of the delivery robot the choice of route or parking spot may have an effect on the length of time it takes to achieve its goal, which in turn may have some financial implications for the company or the customer. Financial implications for the pizza store or the customer could be a consequence that requires evaluation by the robot to make an ethical decision. If we change the example to be that of a medicine delivery robot or an autonomous ambulance robot rather than pizza delivery robot, then the delay in delivery time could lead to more significant undesirable consequences.

8. Consequences Evaluation.

The consequences of each alternative action need to be evaluated. In the case of a game-playing robot, the consequences are evaluated according to a function that involves the likelihood that this action leads to winning. If the selection of an action has an ethical dimension, then the evaluation function would need to compute the utility of each alternative according to utilitarian principles. In the simple case of the delivery robot it could be that each

route could have a different likelihood of causing harm (to itself or to humans) as a result of traffic conditions, the location of schools, and perhaps even prior accident events.

9. Make Decision.

After evaluating the consequences of each alternative action, the robot will select the act that is optimal with respect to the goal or objective of the overall task and optimal with respect to the ethical guidance. In the case of the delivery robot, this step could take place after an evaluation (in step 8) of the relative ethical risks (possible harm) and benefits (e.g. a timely delivery — and hence a reputational and financial benefit to the company). Processing to make decisions could take one of two forms: making decisions after evaluating consequences for alternatives or making decisions that have been made previously and are now located in the machine's training data storage. As mentioned earlier, decision-making must be governed by ethical guidance; therefore, if the decision that has been chosen by the robot is unethical, then the robot could go back to criteria discrimination (stage 6) to choose another ethical alternative.

10. Performing Actions.

This stage results in commitment to the action. The robot may produce outputs, display information on a monitor, activate an actuator, or communicate with another device or person. During this stage the robot should communicate with users to show them that it has taken an action. This means that the output of this stage is external, unlike previous stages which were internal processes inside the robot's system. Communication with users is very important for transparency, as will be illustrated in Chapter 5.

Assessment

The assessment elements govern the learning processes for the training of a robotic system's ethical decision-making. This stage is critical for autonomous self-learning robots. Their learning processes may include an evaluation of the actual consequences of a chosen action compared to the predicted consequences and become input into a data store for training purposes. Such self-learning mechanisms have the virtue of potentially improving the quality of the machines' decision-making processes and increasing the "individuality" of the machines' behaviour: an otherwise identical machine might not behave in the same way under the same circumstances.

This assessment and feedback phase poses an especially important problem from the point of view of allocating responsibilities for action-choices in self-learning machines. An ethically mistaken choice that such an "experienced" machine might make could be the result of prior experiences and any explanation for its behaviour would need to take into account how these experiences have influenced the machine's choice. Therefore, robots that make decisions based on past experiences using machine learning would be using a process which makes it difficult for users or even experts to predict its future actions and explain its previous decisions.

4.1.3.1. Discussion

One virtue of this model is that it provides manufacturers of decision-making machines a framework which defines a system of ethical governance for such decisions. This framework outlines the following: which processes need to be monitored (e.g. to provide explanations in case of accidents); which phases are responsible for consequence-assessments (e.g. when implementing consequentialist ethics); and the role played by deontological "ethical

governance" principles. If, for example, the "generate alternatives" phase fails to consider alternatives that could have avoided an accident, then the question of responsibility for the mistake may reside principally in that module rather than in the prediction or evaluation module. As another example, a system that is self-learning might have failed to generate relevant alternative courses of action because of a fault in the assessment phase. In some situations, failure in alternatives generation may be due to a lack of adequate information collected by the robot's self-learning system during the assessment phase. The robot in that stage should compare between actual action consequences that have been produced by specific actions and the consequences that have been predicted during the predict consequences stage. The assessment phase is essential for a self-learning system because it is responsible for corrections and revisions to improve future decision-making outputs.

The model also provides a framework for legislators and even end-users to understand where ethical responsibilities for machine decisions lie and how to identify the ethical principles that govern these decisions. For example, a machine that is principally driven by a utilitarian calculus for evaluating consequences may be failing to consider alternatives, predicting consequences incorrectly, or evaluating decisions using the wrong objective function. It could be, for instance, that societal values need to be embedded either in the objective function against which consequences are evaluated or in the criteria for discriminating among generated alternatives. Therefore, it is important to take social values into consideration at the time of the robot's design. Note that different societies adhere to different value systems, which need to be respected by robots. For example, personal information privacy varies from one culture to another: a photograph of a woman without a head scarf in the Middle East may not be shared

with others online, and so personal assistance robots in that region must not share users' pictures publicly.

4.1.3.2. Levels of Autonomy

It is important to distinguish between different levels of autonomy, particularly in the decision-making model. Some robots process all stages with no human intervention at all; on the other hand, some robots rely on humans to process a given stage to completion or to proceed from one stage to another. Therefore, we have autonomous robots and dependent robots. For the purpose of this thesis, I distinguish three principal levels of autonomy that decision-making machines can have: low, partial, and full. A robot with low autonomy relies entirely on human input, such as rescue robots or mining robots. Humans are responsible for controlling all of the processing stages described in the model, and in such cases, it is clear where the responsibility for the action lies. Other types of low autonomy robots, such as robot arms used by automobile manufacturers, are machines that require external input at all of their processing stages. They are programmed to run routinely and could not be said to have a meaningful degree of autonomy. A partially autonomous robot only partially relies on human input. These robots have the ability to perform most processing stages by themselves but may need user input to proceed from one stage to the next, such as a human's acceptance of a choice from among alternative decisions. For example, a collaborative robot like Baxter[24], which is used to repeatedly perform only very specific tasks, exhibits some degree of autonomy but does not have the ability to make complex decisions that depend on highly variable environmental conditions; furthermore, it is unable to handle unpredictable situations. Again, in those instances where the ethically critical choice-making step is human-controlled, there is no

[24] http://www.rethinkrobotics.com/baxter/

question about where the allocation of responsibility lies. Finally, a robot is fully autonomous if it relies only on itself, is able to collect and analyze information, generates and evaluates alternatives, and commits to a course of action without human intervention. While these three main levels of autonomy could themselves be subdivided into finer categories, the principal distinctions I make here are sufficient for the purpose and the situations of interest in this thesis where a fully autonomous robot is required to make decisions.

4.1.4. Corporate Social Responsibility

The CSR block in the framework contains transparency and explanation which are required for society to build trust in and reliance upon robots. Consumers must be able to trust the machines that they purchase or live with and be confident that the responsibility for their robots' actions is properly attributed. Clearly, manufacturers should always be held responsible for malfunctions caused by manufacturing defects. However, if a correctly functioning autonomous machine makes a choice that leads to a harmful consequence which in turn brings about a lawsuit, then courts will have to determine who, if anyone, is liable for the harm done.

Laypersons who deal with autonomous robots as users or owners need to know what types of ethical principles are embedded in such robots. This requirement implies that the ethical principles that govern the operations of these robots must be transparent. Additionally, if such robots make mistakes, then users and owners need to know how and why these mistakes happen. This means that the decisions and actions of robots need to be explainable. This need for transparency and explanation is not only for the sake of users and owners but also for the sake of all those who interact with such robots in everyday life.

According to Turilli and Floridi (2009) transparency is "the availability of information, the conditions of its accessibility and how the information, which has been made transparent, may pragmatically or epistemically support the user's decision-making process" and its primary components are accessibility and comprehensibility (p. 106). Thus, a transparent robot "is one in which it is possible to discover how and why the system made a particular decision" (IEEE, 2016, p. 19). Algorithmic transparency in particular "is an emerging research area aimed at explaining decisions made by algorithmic systems" (Datta, Sen, & Zick, 2016, p. 598). Datta et al. define three benefits of algorithmic transparency: (1) identify harms that are caused by algorithms and assign accountability, (2) identify errors in input data that cause inappropriate decisions, and (3) improve the decision-making process. One significant problem with enabling a high level of autonomy in a robotic system is that it increases the opacity of the decision-making process (Zarsky, 2015). A high level of autonomy that includes a self-learning mechanism would lead to opacity in terms of knowing the cause of a robot's decisions. While there is a clear value in enabling autonomous robots with ethical decision-making abilities, it also forces companies that make and program autonomous robots to shoulder social and legal responsibilities and to consider the implications of deploying autonomous robots that are able to make ethical decisions on behalf of users.

The ethical alignment design reports produced by IEEE also advocate for transparency (IEEE, 2016; 2017), for transparency is needed to know how and why a system made a decision or performed an action. They assert that transparency is important for each stakeholder. Transparency is important for users to understand how the system is doing and why it made its decisions, which, in turn, builds trust. Transparency is also important for society to build confidence in intelligent machines. Transparency is also necessary for validation and

certification: it must be possible to disclose a system's inner workings so that it can succeed in the examination process. For accident investigators, transparency is important to understand the causes of failure and for judges, juries, and lawyers to assign legal responsibilities. Based on the IEEE 2017 report, there are four ways in which transparency manifests: as *traceability* which is transparency during implementation of systems for inspection and finding conflict and biases, and as *verifiability* to verify that the system uses norms for decision-making that match and align with the required values. These values include *nondeception*, meaning that the system built is honest in its communication with others; and as *intelligibility* by which the system can explain its behaviours and its reasoning in making decisions or performing actions in a way that is understandable by humans. If the machine cannot explain itself, then it is the designers' responsibility to explain the machine's behaviours.

As I mentioned earlier, as some decisions made by machines become more complex, human owners and consumers will need to know that a machine's decisions are trustworthy and ethically justified. In circumstances where machines make decisions that have undesirable or questionable outcomes, ethical decision-making by autonomous robots requires that their decisions be explainable with human-comprehensible reasons. The problem becomes even more critical as autonomous robots develop greater self-improving and self-learning capabilities and begin making choices and decisions that are based on past experiences. Increasingly, such decision-making scenarios will not be entirely predictable nor clearly explainable after the fact. This combination of non-predictability and autonomy may confer a greater degree of responsibility and accuracy to the machine while at the same time making the machine harder to trust, yet at present it is not normally possibly to explain in non-technical terms why a decision-making machine makes such complex choices. Complex machine

decisions are typically only explainable in complex causal terms — how the algorithm behaved, given the input. Yet to trust the decisions that autonomous robots make, we need to be able to identify which ethical principles — be they consequentialist, deontological, or both — they are applying and how they were applied under the given conditions. Furthermore, humans need to be confident that the robots' application of these principles is deterministic and reliable. As well, this confidence must result from additional components which are illustrated in the Transparency Model in Figure 9.

4.1.4.1. Transparency Model

My transparency model contains three main stages: pre-operation, operation, and post-operation. The first social reason for transparency is that potential robot users need to know the robot's ethical specifications before buying or operating it. I call this the Pre-operation stage. During the operation of a robot users also need to have information about how the robot is functioning. I call this the Operation stage. If an error or problem occurs, then the robot's users need to know how and why the robot made that decision. I call this the Post-operation stage. Table 6 gives a brief explanation of each stage's components.

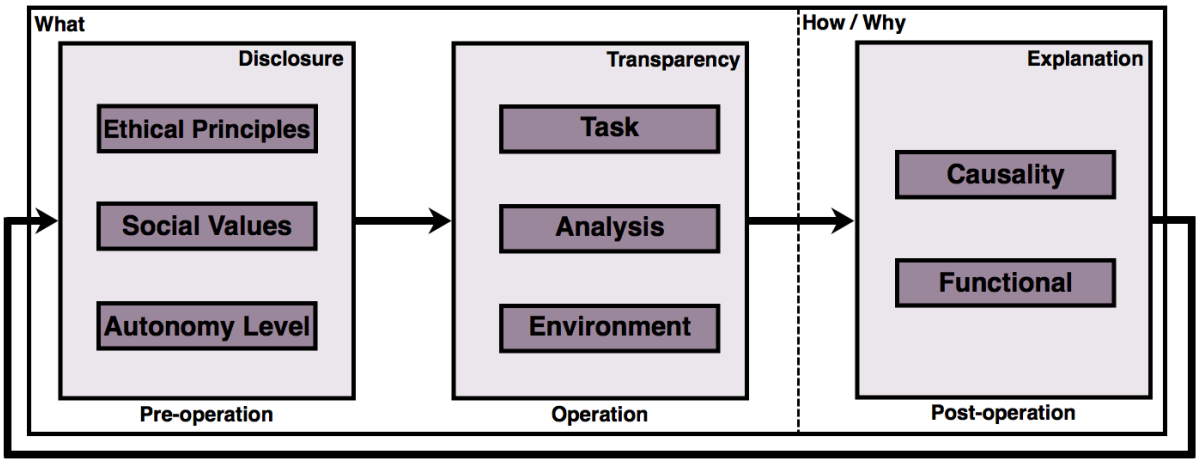

Figure 9 Transparency Model

Table 6 Meaning of Transparency Model's Components

Require	Stage	Purpose	Component	Meaning
What	Pre-operation	Disclosure	Ethical Principles	Embedded ethical principles in the robot: deontological principles, a utilitarian ethical calculus, or a mixture of both.
			Social Values	Social norms with which the robot complies, including regional usage norms, age-ratings, and regional laws.
			Autonomy Levels	Autonomy level either for the robotic system in general or for each decision-making stage.
	Operation	Transparency	Task	Type of task the robot is performing.
			Analysis	Type of information the robot uses in the task.
			Environment	Show and communicate with users that the robot is aware of its surroundings during operation.
How/Why	Post-operation	Explanation	Causality	Declare the cause of problem.
			Functionality	Explain how the robot made a certain decision or performed a specific action.

Pre-Operation

The first stage in the transparency model is pre-operation. The purpose of this stage is for robot corporations to disclose important information for potential users or owners. The pre-operation stage comes before the robot goes into operation — e.g. at the time of purchase — when the users need to know three main components of a robot's specifications, including ethical principles, social values, and autonomy level. The potential users need to know what types of ethical principles are embedded in the robot, be they deontological principles, utilitarian ethical decision procedures, or a mixture of both. The social values component includes what types of social norms the robot complies with, including regional ethical norms, age-ratings, and regional laws. The last component is the autonomy level that the robot manufacturer must reveal about either the robotic system in general or for each decision-making stage. Note that there are other well-known components that could be part of the disclosure stage such as Product Requirement Document (PRD); however, here I am only focusing on the ethical components.

To illustrate the need for disclosure, I use an autonomous vehicle as an example. People will not buy or use an autonomous vehicle if it works mysteriously. There should be a full disclosure of a vehicle's ethical principles (e.g. deontological rules that govern how it reacts to the nearby presence of an emergency vehicle). The vehicle must comply with regional laws such as usage regions (e.g. avoiding some neighbourhoods, see 3.7.4). Finally, a potential user needs to know the vehicle's autonomy level (e.g. level 4) to be aware of his/her role in driving operation.

Labeling Scheme

Labeling, which discloses robot ethical specifications to potential users in the pre-operation stage, is one of the robot corporations' CSR and legal requirements for deploying robots in markets. Just as consumer products have sticker labels (such as the Monroney Sticker or Window Stickers on conventional automobiles), so robot manufacturers will need to declare all of the capabilities of their autonomous robots, including their capabilities to make trustworthy ethical decisions. For example, the labelling of a social autonomous robot used in homes should indicate its embedded social values and ethical principles, the mechanisms that enable it to make ethical decisions, its ability to learn, and its degree of autonomy. Table 7 shows the meaning of each proposed labelling component.

Table 7 Label Components

Label Component	Discloses
Societal Values and Ethical Principles	Ethical principles and social values obeyed by the robot, including its age-rating and the extent to which the robot complies with regional laws, ethics, culture, and values.
Machine Learning and Input Data	Ability of the robot to learn from its experiences, environment, and the user.
System Editing	The extent to which the robot's decision-making system or behaviours can be edited or changed by the user.

Autonomy Level and Role	Level of autonomy for the robot as a whole and for each stage of the decision-making process.
User Privacy and Communication	What type of information it collects and what type of information it shares with other robots, clouds, or the manufacturer.

Operation

The second stage in the model is operational. The purpose of this stage is transparency, where the robot needs to be transparent during functioning. Users and people around the robot need to know what the robot is doing while it makes a decision or performs an action. Joseph Lyons (2013) defines four models for robot-to-human transparency: an intentional model, a task model, an analytical model and an environmental model. For my purposes, I use the task model, the analytical model, and the environmental model. I make some adjustments in these models' functions to align it with the transparency model.

In the task element, the robot should present to the user which type of task it is performing, especially when the robot is multitasking. Additionally, for each subtask, the robot must communicate with the user by providing information about the subtask it is performing and the current situation. The second element is analysis, where the robot presents the type of information it is using in the task that it is performing. This element benefits users by allowing them to understand how the robot makes decisions and performs actions. In the environment element, the robot must show and communicate with users that it is aware of its surroundings during operation. For example, an autonomous vehicle executing a task should be able to show its occupants what type of task it is performing (e.g. slowing down or stopping because pedestrians are crossing the street); what type of information it is analyzing (e.g. distance between vehicles); and what type of object it is detecting (e.g. other vehicles, pedestrians,

cyclists, or road signs). Information shown by the vehicle during operation may differ for occupants and pedestrians, but the main feature is transparency.

Post-Operation

If something goes wrong (e.g. an autonomous vehicle becomes involved in an accident), then society will want to know how and why the accident happened. In this post-operation stage, the ability to provide explanations plays a very important role in enabling acceptance and trust. Explanation is required not just for society — including owners — but also for lawyers, investigators, and judges. Explanations are also needed by government agencies to improve regulations and guidelines that control the autonomous vehicle industry. Additionally, explanation is a very important aspect of allocating responsibility. Companies should therefore be in a position to account for the autonomous actions that they perform, not just with transparency, but also with human-understandable explanations about how and why an event took place when something went wrong. Occasionally, the end-users of autonomous robots will be unable to understand how the machine made its decisions and why they performed specific actions. In some situations, the machines' manufacturers themselves may not be able to understand why their product made some decisions, either because the machine learning method used for developing such a capability cannot provide human-understandable explanations or because of the complexity in the machines' algorithmic decision-making process. Such inexplicable processes are beginning to manifest in decision-making machines that rely on machine learning. For example, when DeepMind's AlphaGo played against Lee Sedol in March 2016, it made unusual moves that were unexpected by its designers and not easily understandable at that time, either by professional Go players or by the software designers (Moyer, 2016). Therefore, if an autonomous robot makes a decision or performs an

action that either is or appears to be ethically mistaken, then the manufacturers should be able to explain to users how and why these decisions were made. In addition, manufacturers need to be able to tell users what they need to do to improve their robots' decision-making systems to protect themselves and others from future failures.

For example, on March 18, 2018 an autonomous vehicle operated by Uber struck and killed a woman who was crossing the street in Tempe, Arizona[25]. Although the vehicle was equipped with sensors and systems for autonomous driving, these technologies did not prevent this shocking accident. A few days after the accident the authorities released a short video from the vehicle's dash camera that shows what happened before the accident. An answer to the question of *how* the accident happened was provided by *The New York Times*[26]. Answering the first question of *how* is equally important as addressing the question *why*, which has not yet been answered.

In my transparency model, there are two principle elements in an explanation — causality and functionality. This explanation stage is the robot corporations' responsibility and comes post operation. In the causality element, the robot's manufacturer must be able to determine the cause of the problem, which may be the absence of ethical principles in general, the absence of social values, a technical malfunction or a mistake in the ethical decision makine process. The robot's users, in turn, need to know the cause, as do those who live with or around that robot. The causality element answers the question *why* and functionality element answers the question *how*. Therefore, in the functionality element, the robot manufacturer needs to be able

[25] Wakabayashi, D. (2018, March 19). Self-Driving Uber Car Kills Pedestrian in Arizona, Where Robots Roam. Retrieved March 28, 2018, from https://www.nytimes.com/2018/03/19/technology/uber-driverless-fatality.html
[26] Griggs, T., & Wakabayashi, D. (2018, March 21). How a Self-Driving Uber Killed a Pedestrian in Arizona. Retrieved March 28, 2018, from https://www.nytimes.com/interactive/2018/03/20/us/self-driving-uber-pedestrian-killed.html

to explain, in detail and in terms that are understandable by ordinary people how the robot made a certain decision or performed a specific action. Wolter Pieters (2010) defines four levels of explanation based on the level of detail in information that should be given to users. The explanation will fail if there is not enough detail, and the questions of *how* and *why* are not answered (superficial reasons). The explanation will fail if too many details are given that are only understood by experts in the field (deep details). Therefore, to reach a suitable and understandable explanation for ordinary people, the level of detail should adequately explain why and how. Explaining why gains users' confidence and explaining how gains users' trust.

Generally, transparency and explanation in the trust framework enhance acceptance and trust of robots among members of society. As I mentioned earlier, the transparency model, including the disclosure, transparency, and explanation components is useful for validation and certification and examination as well as to investigators, judges, juries, and lawyers. In addition to those benefits, the trust that is engendered by transparency will transfer to potential investors in the autonomous robotics industry. Robots will not last the marketplace if they do not gain acceptance and trust by users and citizens in society.

4.1.5. Sustainability

The sustainability block in the framework (Figure 7) contains the trust and reputation elements. Trust, which is the result of transparency and explanation, affects corporations' reputations which in turn impacts the sustainability of corporations in the robot industry. Robot companies must maintain their reputation by reinforcing society's trust in their products. The sustainability process is ongoing with all new robot products and new versions of existing robots.

Since future robots will have a greater capacity to perform an increasing number and variety of tasks, including some with ethical consequences, robots must not only follow the ethical rules that they are programmed to obey but also be seen to be following these rules explicitly. They must be trustworthy and perceived as trustworthy. Trust has multidisciplinary definitions based on different concepts and relationships (Cho, Swami, & Chen, 2011). For example, trust can exist between one individual and another, between systems in a network, and between a user and an intelligent agent. Kirkpatrick et al. explain: "Trust can take various forms; it can range from institutional trust … to trust in oneself … to trust in one's government …" (Kirkpatrick et al., 2017). Therefore, while trust could take different forms, the trust to which I refer in this framework is divided into two types: (1) the set of relationships between users and robots but also between members of society as a whole and the robots upon which they rely, and (2) the relationships between society and corporations that create robots. Therefore, for corporations to be permitted to deploy their robots in society, governments must ascertain that the robots are trustworthy enough to be deployed. Users must be able to trust robots to decide and act on their behalf, users must trust the corporations that create them, society must trust the governments that regulate robotics systems, and people must trust the robots that live with them. Embedding explicit and transparent ethical principles in robot systems would enable trust. Therefore, if everyone in society, including individual users, know and observe that robots have such ethical principles, then we could become confident that these principles are protecting us from potentially harmful decisions. For instance, in the Amazon example given previously where Amazon's recommender recommended to users who had added a harmless substance to their shopping basket the remaining ingredients that together make a bomb (see Algorithmic Bias), one way to prevent users from unethical emergent bias would

be to formulate a new version of Asimov's First Law of Robotics[27] which states: 'Do not recommend dangerous products that could harm them or others or recommend products that are dangerous if they are combined together'.

Regarding the reputation component in the trust framework, the robot company's reputation is gained over time by stakeholders' evaluations, based on stakeholders' experiences (Gotsi & Wilson, 2001). Company stakeholders' experiences rely on the impacts of three factors: financial, social, and environmental (Barnett, Jermier, & Lafferty, 2006). In general, reputation can be thought of as the collective opinions of users or members of society about the behaviour of someone or something based on their past actions (Granatyr et al., 2015). The difference between trust and reputation is that "Trust [is] a peer's belief in another peer's capabilities, honesty and reliability based on its own direct experiences; [while] Reputation [is] a peer's belief in another peer's capabilities, honesty and reliability based on recommendations received from other peers" (Wang & Vassileva, 2003, p. 150). Therefore, potential users of a product may or may not buy the product based on its reputation and its creator's reputation. That positive or negative reputation comes from users' experiences with the company or the product. There are many benefits of maintaining a corporate reputation such as attracting investors to the industry, raising the corporate market value, and attaining better market position (Feldman, Bahamonde, & Velasquez Bellido, 2014). These benefits have positive impacts not only on the robot manufacturers themselves but also on the robotics industry in general. Therefore, gaining users' trust has a positive effect on corporate reputation.

[27] "A robot may not injure a human being or, through inaction, allow a human being to come to harm."

In the business world, corporate reputation is important. In the context of the trust framework, reputation is the element that depends on trust, which is a very significant factor for sustainability. People in the near future will rely on autonomous robots to make their daily lives easier; hence, products that cause harm could have dramatic effects on corporate reputations and devastating effects on users' trust. For example, corporate reputations could be damaged to the point of affecting sales and corporate stock valuations, as happened in the aftermath of the Volkswagen emissions scandal (Zhang, 2016). Just as Volkswagen's reputation has been affected by its rigging emissions testing software, so too the potential for this kind of reputation loss could happen to a robotics company if even one of its autonomous robot makes a mistake. For example, the reputation of Uber could be damaged after the fatal crash involving one of their autonomous vehicles. For corporations to maintain their reputations, their robots must be reliable, safe, transparent, respectful of laws and values, and trustable. Robots that have the ability to make decisions and perform actions autonomously will have an observable impact on the reputations of the corporations that manufacture them.

In summary, this chapter answers my research questions as shown in Table 22. Generally, the answers show the relationships among different components, namely stakeholders, robots, computing, ethics, CSR, and sustainability. The interactions between these factors demonstrate four important relationiships: 1) The role of regulators in the robotics industry is to guide companies to build ethical robots and to protect society from robots' failures; 2) The role of ethical principles and social values is to control robots' behaviours through ethical decision-making processes; 3) The role of disclosure, transparency and explanation is to make societies accept, trust, and rely on robots to make decisions; and 4) The role of trust is to maintain the sustainability of robotics companies' reputations.

5. Content Analysis

To obtain corroborating evidence for the theoretical transparency model proposed in the previous chapter, I chose to study public opinions to better understand what laypeople think about social autonomous robots that have the ability to make decisions that have ethical consequences. To conduct such a study, I needed to locate a forum where people share their ideas and thoughts about such robots. However, at present there are no commercial fully autonomous social robots on the market that are being used by the public. Therefore, the majority of people do not have experience interacting with fully autonomous robots. Fully autonomous robots are fully independent (see Table 3 and Table 4), make decisions autonomously, and perform actions without any human intervention. It is challenging to study opinions from members of the public about something that does not yet exist. Still, autonomous vehicles have received a lot of attention recently for various reasons. For example, automotive and technology companies have competed to build such vehicles, media outlets have given some attention to this technology, and academic scholars have shared their thoughts either in research articles or on websites. In addition, there are semi-autonomous vehicles (level 2 and level 3 in Table 5) that are now being used by people on public roads; therefore, some people have begun to form an idea about their functions and abilities.

Consequently, the discussion forum I chose as a source of data (opinions or comments) needed to have been expressed by the public on autonomous vehicles. Hence, I tried to find an article or a study that discussed autonomous vehicles, had been published publicly (on a website), and included a comment section where readers can share their thoughts. Note that autonomous vehicle articles and research studies discuss many related issues such as the expected and

potential benefits, impacts on the economy, society, environment and public health, safety and security, sensors and communications, trust, bullying, decision-making, liability and responsibility, and insurance. Thus, I decided to choose an article or a research study that focuses on ethical decision-making to answer two of my research questions: (Q2) If humans are to trust robots to make ethical decisions, what are the conditions under which this trust is justified? (Q4) How do ethical decision-making algorithms need to be constrained in order to make them acceptable to the public? To be more specific, my purpose in studying the public's opinions was to examine the following: (1) what people think about ethical decision-making in social autonomous robots, (2) how people accept such robots, (3) whether people will trust these robots, and (4) whether people prefer utilitarian or deontological ethics in autonomous robots that face trolley problem situations.

One of the most discussed issues in ethical decision-making in autonomous vehicles is identifying the right decision in an unavoidable crash. In fact, several articles discuss how autonomous vehicles make decisions during crashes and how autonomous vehicles are programmed to reduce damage and harm, which is the modern version of the trolley problem. Therefore, as illustrated in Table 8, among eight sources/articles that discuss ethical decision-making in autonomous vehicles, I chose two sources (article number 4 and article number 6) based on the number of comments.

Table 8 List of Comment Sources

N	Title	URL	Discussion On	# of Comments
1	Self-Driving Cars Will Kill People. Who Decides Who Dies?	https://goo.gl/8Qe4CA	Wired.com	102
2	How to Help Self-Driving Cars Make Ethical Decisions	https://goo.gl/N8HqBL	technologyreview.com	34
3	To Make Us All Safer, Robocars Will Sometimes Have to Kill	https://goo.gl/51EQhY	wired.com	37
4	Moral Machine	https://goo.gl/W7GZ2P	moralmachine.mit.edu	374
5	The Ethics of Autonomous Cars	https://goo.gl/uvWBBk	theatlantic.com	137
6	Why Self-Driving Cars Must Be Programmed to Kill	https://goo.gl/6JofJi	technologyreview.com	371
7	Self-Driving Cars Pose Thorny Ethical Questions in AI	https://goo.gl/m9h1fR	extremetech.com	39
8	Can You Program Ethics into a Self-Driving Car?	https://goo.gl/PMc9fp	spectrum.ieee.org	16

Regarding the first source (henceforth "The Game"), the significance of the trolley problem for machine ethics and autonomous vehicles in particular is illustrated well by the thought-experiments shown on the Moral Machine Game webpage (http://moralmachine.mit.edu/). This is one of the most well-known web platforms for experimenting with different ethical choice scenarios. By posing moral dilemmas that could be faced by autonomous vehicles in unavoidable crash situations, these scenarios are presented to users for ethical evaluation. In each scenario (thirteen scenarios in total), users need to choose between two harms based on their ethical preferences. Figure 10 shows one of The Game's scenarios. I also chose this source because the web site developers collected thirty million decisions by more than three million users from 160 countries in the period between its deployment on June 23rd, 2016 until May 2017 (Awad, 2017). Awad (2017) observes that these webpage developers have in fact collected the largest dataset in the history of artificial intelligence ethics. Therefore, this well-known game encourages visitors to participate to see how their morals control their decisions. For the purpose of this thesis, I wanted to study The Game participants' thoughts and opinions

about the nature of The Game and their feelings toward autonomous vehicles that make ethical decisions. Thus, I focused on the comments section on the webpage as a source (data corpus) to collect public opinions.

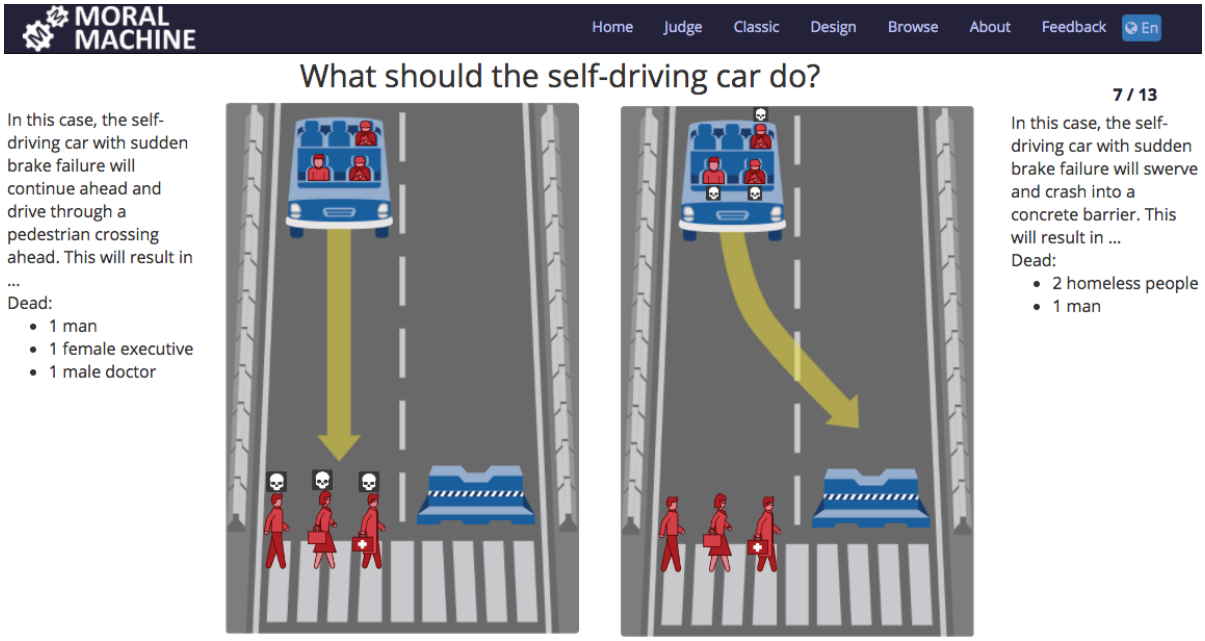

Figure 10 Moral Machine Game

To increase the variety of points of view from by members of the public about the same issue of ethical decisions made by autonomous vehicles - I selected another source (henceforth "The Article"): comments on an article entitled "Why Self-Driving Cars Must Be Programmed to Kill" published in *MIT Technology Review*. I also chose this article because it discusses the acceptance of autonomous vehicles which make ethical decisions. At the same time The Article reviews a portion of the research by (Bonnefon et al., 2015) which focuses on how people accept and trust utilitarian vehicles. Commenters from each source shared their thoughts about ethical autonomous vehicle decision-making in unavoidable crash situations.

The developers of both sources use a comments plug-in to encourage their audiences to share their ideas and become engaged in the discussion. The plug-in, produced by Disqus.com, is one of the most widely used comment plug-ins for publishers, because it allows readers who have a Disqus account to post comments on any website that uses its plug-in. In addition, readers have the option to post their comments on Disqus by using their Facebook, Twitter, or Google+ accounts or their own universal Disqus accounts. Finally, any comment posted by a user on any website that uses Disqus plug-in will be saved in the user's profile on Disqus.com.

For a researcher, this feature of the Disqus plug-in is useful for two reasons. First, it is used by several websites that publish articles on websites which discuss different issues about autonomous vehicles including 9to5google.com, extremetech.com, spectrum.ieee.org, theatlantic.com, technologyreview.com, and wired.com. In addition, the Disqus plug-in enabled me to save comments in bulk by clicking on the Recommend button that appears in any article to my profile on Disqus.com, hence making it easy for me to save several sets of comments for the purpose of study and comparison. While there are other commenting platforms whose contents could also be harvested, Disqus is the most widespread and the content it contains suffices to provide the evidence I need in support of my hypotheses concerning the Transparency model.

Table 9 shows my data set containing two sets of comments: one set on The Game and one set on The Article. It is important to note that these comments were all posted on the public internet and that no research ethics application was required.

Table 9 Data Set Sources

Data Type	About	Title	Source Link	Comments Link
Comments	The game	Moral Machine	https://goo.gl/W7GZ2P	https://goo.gl/rj6tb2 https://goo.gl/yajkZs
Comments	The article	Why Self-Driving Cars Must Be Programmed to Kill	https://goo.gl/6JofJi	https://goo.gl/Wpk3gj

5.1. Data Set and Method

From among the 785 comments harvested from these sources, many occurred in nested discussion threads. Of these comments in discussion threads, some digressed from the original topic. Therefore, I used the following selection criteria: (1) The main criterion was to choose comments that were posted by the first author only, so I distinguished between initial comments and replying comments. I disregarded all replying comments because some comments led the discussion away from my scope. Figure 11 shows part of a discussion thread in which commenters started by discussing machine learning and ended up discussing racism. (2) I ignored all useless comments such as sarcastic, mocking, and spam comments. In any case, these were few.

Therefore, out of 785 comments from both sources I selected 287 initial comments for analysis. This produced 496 opinions. Table 10 provides more details about the data sets. In this context, an "opinion" is the occurrence of a statement in a comment that expresses an individual thought, as identified by a coded category which is discussed in more detail in the Coding Process section. Using several coding methodologies, the third comment in Table 13 was identified as expressing (1.2, 3.1, and 4.1) opinions. Each comment in The Game has, on average, 1.8 opinions. Comments in The Article have on average, 1.5 opinions per comment.

Table 10 Number of Comments, Opinions, and Ratio

Title	# of Comments	Initial Comments	Opinions	Ratio
Moral Machines	374	115	225	1.8
	40	12		
Why Self-Driving Cars Must Be Programmed to Kill	371	160	241	1.5
Total	785	287	496	

Figure 11 Sample of DISQUS Comments Leading to Digressions

Knowing that there are similarities between the two sets, it is important to analyze the two sets of comments using two different sets of categories and codes because the comments represent different thoughts that have been discussed/presented in the source text. For each scenario in the Moral Machine Game, users are asked to choose between two decisions. In all scenarios the users must choose to sacrifice either vehicle passengers or pedestrians in an unavoidable crash due to a sudden brake failure. Note that in all scenarios there are different types of potential victims represented by age, weight, sex, and social position such as doctor, athlete, homeless, executive, and criminal, clearly not categories that one would expect an autonomous vehicle to ever be able to (or allowed to) distinguish. Also, in some scenarios there are also animals among the victims and in some scenarios, pedestrians are flouting the law by crossing the street while the pedestrian light is red. The main purpose of The Game is to make users choose who should be sacrificed and who should survive in different circumstances.

On the other hand, The Article uses the same ethical decision-making dilemmas faced by an autonomous vehicle during an unavoidable crash. The Article discusses the acceptance of autonomous vehicles, which, as mentioned previously, is one result of (Bonnefon et al., 2015) research. After carefully reading the comments in each of The Game and The Article, I came up with a coding system for each set of comments by following and using methods from (Saldana, 2009; 2015) as described below.

5.2. Coding Process

Since a significant proportion of comments expresses several opinions, and since many of these opinions are repeated by participants, I needed a method to highlight, extract, and count these comments. For this purpose, I chose magnitude coding, which consists of adding a word

or a phrase that indicates the presence of an opinion in the data that enables the researcher to count the frequency of its occurrences and its percentages in each category (Saldana, 2015). For example, there are many commenters in both sets that rejected the idea of machines that make ethical decisions; therefore, I coded this opinion as *Machines Shouldn't Decide*.

The coding emerged from reading and analyzing the data sets several times. Before the analysis, I did not have any preconceptions about the themes that would arise. To start coding, I followed several steps for each data set separately. (1) I copied all comments that I had chosen and pasted them in a Word document. (2) I completed the first cycle of coding by summarizing each comment separately using simple words. I made sure that the summary would contain the major ideas of the comment as preliminary codes. (3) After I finished summarizing all of the comments, I repeated the first cycle by examining the summaries to find similarities and frequencies of ideas and determined initial codes. (4) I started the second cycle of coding by sorting and classifying the *major ideas* that are repeated in all comments to identify the main categories. (5) I studied each comment again to understand the reason *why* the commenter chose to voice a specific idea. For example, in The Game's set of comments some commenters chose to sacrifice pedestrians because they were jaywalkers. Therefore, the category for that comment is *Sacrificing Pedestrians* and the code for the reason is *Break the Law*. Appendix A and Appendix B show the chosen comments, summaries, and their codes.

Some comments contain more than one major idea. These comments which fit into different categories and reasons (codes) could be called simultaneous coding. In the same context, some comments have more than one reason in one category. Using the previous example, some commenters chose to sacrifice pedestrians because they were jaywalkers, and at the same time, these commenters reasoned that pedestrians would have a chance to survive by running away

from the malfunctioning vehicle once they saw it coming toward them. The category for that comment is *Sacrificing Pedestrians* and it has two codes in one category for two reasons: namely *Break the Law* and *Avoiding Vehicle*. Figure 12 shows the code model.

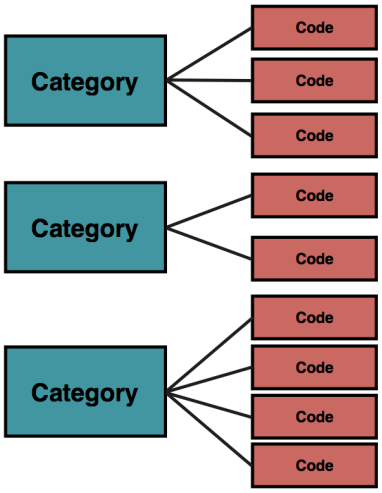

Figure 12 Coding Model

5.3. The Game's Comment Analysis

In this section, I explain how I came up with the categories and reasons for The Game comments. I divided the comments based on commenters' preferences about who should be sacrificed into four categories: *Sacrifice Passengers, Sacrifice Pedestrians, Alternative Solutions, and Refraining*. Some commenters chose to *sacrifice vehicle's passengers* while others chose to *sacrifice the pedestrians* and each group supported their choice with varying arguments. In addition, some commenters did not choose to sacrifice either passengers or pedestrians but instead suggested *alternative solutions* to solve the problem of an unavoidable crash situation. The last group of commenters did not choose to sacrifice anyone and also did not provide any alternative solution, and so I called their comments *Refraining*. As I have mentioned, there are different reasons which motivated commenters when selecting who to

sacrifice, suggesting solutions, or refraining. For that reason, I divided each of the four categories into sub-categories according to the reasons that were offered, summarised in Table 11.

Table 11 Categories of the Game Comments

Category	Reason (Code)
Sacrifice Passengers	No Maintenance
	Their Choice to Ride
	Cars are Safe
Sacrifice Pedestrians	Break the Law
	Avoid Vehicle
Alternative Solutions	Swerve
	Keep Straight
	Reduce Deaths
	Disable the Vehicle
Refraining	The Dilemma is Unrealistic
	Machines Shouldn't Decide
	No Trust
	Greatest Happiness
	Other

The commenters who support the idea of *sacrificing passengers* had three different reasons. Some commenters sacrificed passengers because they had failed to do *Maintenance* on their vehicle, which was the cause of the vehicle's brake failure. Other commenters sacrificed passengers because *They Chose to Ride* such vehicles and therefore they should be responsible for the consequences of their choice. The last group who sacrificed passengers said that vehicles are equipped with many safety features and the passengers would have a high probability of surviving if the vehicle swerves to hit a wall or a barrier; therefore, *Cars are Safe*. In contrast, the pedestrians are not protected by safety features and so they will likely die or be severely injured if they are hit by a car. Therefore, the first category of *Sacrifice Passengers* has three reasons.

In the second category, commenters supported the idea of *Sacrificing Pedestrians* for one of two reasons. First, pedestrians who cross the street while the pedestrian light is red are jaywalkers — *Break the Law* — and they should be sacrificed due to their carelessness regardless of their number or position. Second, pedestrians have the opportunity to run to *Avoid the Vehicle*, and likely avoid the crash before it happens. Thus, the second category of *Sacrificing Pedestrians* has two reasons.

The third category is comprised of commenters who did not — or did not want to — declare their choice but instead suggested some *Alternative Solutions*. Thus, this category is divided into four reasons. One group of commenters said that the vehicle should *Swerve* to hit a wall or a barrier no matter who and how many are in the vehicle or on the street. Another group of commenters proposed another solution which is that the vehicle should *Keep Driving Straight*. The vehicle should not change its lane because changing lanes would produce other problems such as hitting other vehicles or running over pedestrians on the sidewalk, thereby increasing the number of victims. In addition, a number of these commenters suggested that in addition to keeping straight, the vehicle should warn and alert pedestrians on the street for example by sounding an alarm to give them chance to run. The third reason was given by a group of commenters who said that *Reducing Deaths* was the most important goal, regardless of how the vehicle does this. Finally, some commenters suggested that the *Vehicle should be Disabled* with emergency features such as a double emergency brake or shutting down the engine and using an active gear emergency brake to stop the vehicle immediately. In total, the third category of *Alternative Solutions* has four reasons.

The fourth category is *Refraining* comments, which did not express thoughts of sacrificing either passengers or pedestrians nor provide solutions. This category was justified with five

reasons. The first group includes commenters who refused the premise of The Game and said that the *Dilemma is Unrealistic*. These commenters claimed that the circumstances depicted in The Game are not possible in reality and would not happen as described in The Game. Also, these commenters said that modern vehicles will be developed to the same standards as almost all other vehicles on the road, and so the brakes would not fail. The second group affirmed that *Machines Should Not Decide* and choose who to kill and who should survive. The third group stated that they *Do Not Trust* autonomous vehicles and they would not buy them in the future. The fourth group is comprised of commenters who confirmed that although there will be such dilemmas in the future, in comparison with the current situation, autonomous vehicles will be better than humans at driving because in general they will reduce the overall number of accidents and deaths. Considering that these commenters reflect a utilitarian perspective from the point of view of society as a whole, I coded this reason as *Greatest Happiness*. The last group is comprised of commenters who shared *Other* useful ideas that are uncategorized, such as the idea that machine learning has the ability to recognize people's ages and genders with high accuracy. Thus, the fourth category of *Refraining* has five reasons.

5.3.1. Recoding The Game Comments

To extract the frequency and patterns of the thoughts and ideas and to discover the most/least repeated reasons and the correlations among them, I needed to calculate frequencies and percentages as well as analyze all the codes. Calculating all categories and reasons without using software required a transition phase of recoding. Therefore, I decided to use MS Excel to perform the calculations. For this task I redefined each category and reason using numerical symbols for speed, ease, and conciseness. For the purpose of clarity, I used *natural numbers* from 1 to 4 for each category (since I have four categories). For each reason I used *rational*

numbers from 1.1 to 4.5, the first digit of which indicates the category number, and the second digit indicates the reason number as shown in Table 12. Consequently, I returned to the second cycle of coding (steps 4 and 5 in the Coding Process section) and beside each comment I replaced each category and reason with the assigned number. For example, beside the comment that has the idea of *Sacrificing pedestrians* because they were jaywalkers, I wrote number 2 for the category of *Sacrificing pedestrians* and 2.1 for the reason of *Break the Law*. After recoding all comments by following this process, I obtained a whole set of comments, summaries, category natural numbers, and reason rational numbers as shown by the sample in Table 13 (all comments found in Appendix A).

Table 12 The Game's Numerical Symbols

Category	Numerical Symbols	Reason	Numerical Symbols
Sacrifice Passengers	1	No Maintenance	1.1
		Their Choice to Ride	1.2
		Cars are Safe	1.3
Sacrifice Pedestrians	2	Break the Law	2.1
		Avoid Vehicle	2.2
Alternative Solutions	3	Swerve	3.1
		Keep Straight	3.2
		Reduce Deaths	3.3
		Disable the Vehicle	3.4
Refraining	4	The Dilemma is Unrealistic	4.1
		Machines Shouldn't Decide	4.2
		No Trust	4.3
		Greatest Happiness	4.4
		Other	4.5

Table 13 The Game's Recoding Sample

N	Comment	Summary	Category	Reason
1	I would program the car to stop sideways against a wall. use a wall as a break.	Hit a wall to stop	3	3.1
2	Seriously, when was the last time a modern car just lost all braking ability?	Unrealistic	4	4.1
3	The car is at fault; the people in the car should bear the responsibility for the fault. In every case, I'd rather the car hit the wall than hit a pedestrian. This is a stupid test in any case. So many ways in practice to avoid this dilemma.	Passengers' responsibility Hit a wall Unrealistic	1 3 4	1.2 3.1 4.1
4	…	…	…	…

5.3.2. Calculating The Game Comments

To obtain the results listed in Table 14, I copied the last two columns — Category and Reason — from Table 13 and pasted them in Excel. (1) The first step is counts how often a single category occurs by using COUNTIF function (e.g. to count the first category which is *Sacrifice Passengers* I used the formula = COUNTIF(A2:A226,1)), and then counted its percentage. (2) The second step counts how often a single reason occurs by using the same function and then counts its percentage to calculate how much each reason weighs in each category. (3) The third step counts the percentages of each reason out of the total number of reasons to identify the most frequently repeated reasons in all comments.

5.3.3. Results of The Game Analysis

Table 14 The Game's Calculation Results

Category	Frequency	%	Reason	Code	Frequency	%	% out of 225
Sacrifice Passengers	21	9.3%	No Maintenance	1.1	3	14.3%	1.3%
			Their Choice to Ride	1.2	16	76.2%	7.1%
			Cars are Safe	1.3	2	9.5%	0.9%
Sacrifice Pedestrians	22	9.8%	Break the Law	2.1	16	72.7%	7.1%
			Avoid Vehicle	2.2	6	27.3%	2.7%
Alternative Solutions	87	38.7%	Swerve	3.1	15	17.2%	6.7%
			Keep Straight	3.2	21	24.1%	9.3%
			Reduce Deaths	3.3	36	41.4%	16%
			Disable the Vehicle	3.4	15	17.2%	6.7%
Refraining	95	42.2%	The Dilemma is Unrealistic	4.1	19	20%	8.4%
			Machines Shouldn't Decide	4.2	30	31.6%	13.3%
			No Trust	4.3	15	15.8%	6.7%
			Greatest Happiness	4.4	11	11.8%	4.9%
			Other	4.5	20	21.1%	8.9%
Total	225	100%			225		100%

Table 14 shows the summary results of the comment analysis. Only 19.1% (9.3% + 9.8%) of comments supported the idea of *Sacrificing* either *Passengers* or *Pedestrians*. In terms of *Sacrificing Passengers,* 14.3% said that passengers must not sacrifice anybody except

themselves because they did not maintain their vehicle and this lack of vehicle maintenance led to brake failure. For instance, in comment number 1 (see the Appendix A) the author says, *"All of these scenarios assume sudden brake failure so in most cases I would say it's the car owners fault for driving a non maintained vehicle and [the car owner] should be the one to suffer as a result."* 76.2% of commenters who chose to *Sacrifice Passengers* did so because the passengers were the main reason that the vehicle was on the road; thus, the passengers must be ready to face the consequences of their choice. For instance, in comment number 3 the author says, *"I think that if you make the decision to let the car drive you around, you should always die rather than kill someone else."* Therefore, nobody should be sacrificed in any situation other than passengers. 9.5% of commenters said that since modern vehicles have advanced safety features, *Sacrificing Passengers* is the best choice because they have a high likelihood of surviving. Comment number 21 illustrates this point of view well: *"I think the passengers should Always take the risk because hitting a pedestrian with a vehicle, that pedestrian has No chance! But if you're strapped in with a safety belt, with air bags and engine crumple zones, etc etc etc, vehicles have a million safety features to protect passengers, but 0 to protect pedestrians."*

On the other hand, 72.7% of commenters supported the idea of *Sacrificing Pedestrians* because pedestrians were the ones flouting the law. For example, the author of comment number 24 said, *"People that jay walk know the risk/consequences and should pay for it."* Therefore, it is unfair to *Sacrifice Passengers* because of jaywalkers who broke the law by crossing the street. 27.3% of commenters said that pedestrians could avoid the accident and survive. For example, in comment number 117 the author reasons, *"continuing straight will allow the pedestrians to move out of the way."*

Although the premise of The Game is that the user must select a preference for the autonomous vehicle to sacrifice either passengers or pedestrians, the majority of comments (80.9%) either suggest *Alternative Solutions* or *Refraining*, neither of which involves making a sacrifice. Those commenters had a variety of reasons for not wanting to decide who should die. 38.7% suggested some *Alternative Solutions* to reduce or avoid any harm. 42.2% were *Refraining* from participating in sacrificing or suggesting alternative solutions. The majority of these commenters expressed a fear of autonomous vehicles. There are four suggestions in the *Alternative Solutions* category: *swerve, keep straight, reduce deaths, and disable the vehicle*. 17.2% suggested that the vehicle should *Swerve* and hit a wall or barrier to stop itself from moving, thereby preventing further damage. Keeping straight and hitting pedestrians does not mean that the vehicle will stop. Instead, it may run over pedestrians, continue driving, and crash into more vehicles or hit other pedestrians on the other side of the intersection. For instance, in comment number 51 the author says, *"i[n] all cases involving brake failure and concrete block i go for the concrete block , elsewise the car will continue its course and prolly kill more ppl. doesnt really matter who is in the car or who is walking."* On the other hand, 24.1% of commenters recommended that the vehicle should *Keep Driving Straight* because swerving could amplify the damage by hitting other pedestrians on the sidewalks or hitting other vehicles in other lanes. For example, the author of comment number 117 said that *"Swerving into a pedestrian who has his back towards the car decreases his chances of getting out of the way."* In addition, some commenters suggested that the vehicle should drive straight and warn pedestrians by honking or sounding an alarm to give the pedestrians a chance to move off of the street. For instance, in comment number 66 the author says, *"At least: honk!"* Programming such vehicles to drive straight and sound an alarm in unavoidable crash situations would allow for transparency to the public who use and live with those machines;

in addition, that action would be predictable by people. For example, in comment number 13 the author says, *"I feel the car should always move in a predictable direction, ie stay in lane, regardless of who is being driven or who is on the crossing."*

The third solution is *Reduce Deaths* by any means because humans are a priority no matter who they are. This observation was made by 41.4% of commenters. In fact, *Reducing Deaths* is the most frequently chosen reason among all reasons (mentioned in 36 out of 225 opinions, or 16%). This reason is actually a manifestation of utilitarian thinking. Hence, if the killing is unavoidable, then one should at least kill the fewest people. For instance, the author of comment number 82 says, *"In any other case I think the only way to go is the one that keeps the highest chance for people (no matter who, no matter if inside/outside the car) to survive."* The last solution of *Disable the Vehicle* was expressed 17.2% of the time. Commenters also suggested the addition of new safety features that protect passengers and pedestrians. For instance, in comment number 37 the author says, *"Total brake failure can (and should) be mitigated by aggressive downshifting and stopping the engine. Can also place car in park or reverse. This may result in damage or complete destruction of the engine/transmission/car, but that is better than killing someone."* The author of comment number 54 says, *"the car really just needs a parachute or eject wheels button though"*, and in comment number 86 the author says, *"let's think how we can stop the car if primary brakes fails and forget about killing people, let's add airbags to the front of the car, since the car can now predict the accident and identify the surrounding, it can open the airbags to reduce the number of casualties."*

Refraining comments occurred 42.2% of the time. The commenters did not choose to sacrifice or suggest solutions for four reasons: (i) *The Dilemma is Unrealistic*; (ii) *Machines Should Not Decide* who dies; (iii) *No Trust* in such vehicles; and (iv) the greater social good (*Greatest*

Happiness of the greatest number). Twenty per cent of commenters who abstained rejected the idea of The Game because it is *Unrealistic*. For instance, in comment number 6 the author concludes, *"This is all silly. We don't morally or legally require human drivers to perform stunts to avoid collisions of any kind."* The author of comment number 23 says, *"Such a gruesome test. I felt dirty just taking it."* In comment number 59 the author asks, *"Seriously, when was the last time a modern car just lost all braking ability?"* The next reason suggested is *Machines Should Not Decide* who dies (31.6%). For example, in comment number 15 the author says, *"It's pretty simple really, a car shouldn't have the ability to judge who it's about to run over."* In comment number 52 the author says, *"You cannot let a machine be judge and jury. You cannot let a car decide the value of other human beings based on outside factors."*

The third reason given in the *Refraining* comments is the absence of trust (*No Trust*) (15.8%). Some commenters said that they would not buy a vehicle that sacrifices its passengers; for instance, in comment number 79 the author says, *"I would never buy a self driving car if I knew there was a POSSIBILITY it would ethically choose between killing me or other people."* Actually, the last two reasons — *Machines should not decide* and *No Trust* — could be collapsed into one reason of unacceptability. The commenters who chose these two reasons did not accept the idea of an ethically autonomous vehicle that could decide who would die. Thus, if the number of their occurrences are added to one other (30 + 15), then they are the most frequently stated reason representing 20% out of 225 opinions. Therefore, one-fifth of participants did not accept a vehicle that could sacrifice its passengers or decide to kill pedestrians.

The fourth reason for *Refraining* was given by commenters who chose the greater social good — *Greatest Happiness* of the greatest number (11.6%). Although some commenters did not

accept the idea of The Game, they were sure that technology would save people's lives. For example, in comment number 14 the author says, *"It's easy to come up with hypothetical scenarios like this, but we should not loose (sic) the focus of how much good a switch to self driving cars would bring in terms of the overall accident (fatality) rate in traffic."*

The last reason in the *Refraining* category is *Other*, which I used to classify comments that do not contain any *Sacrificing*, *Alternative Solution*, or *Refraining* but present useful ideas. These comments represent 21.1% of *Refraining* comments. For example, in comment number 12 the author says, *"A society capable of developing a system of driverless cars is also capable of permanently separating vehicles from any area used by pedestrians."* Thus, this comment supported the idea of building designated areas for autonomous vehicles that are not allowed for pedestrians. Indeed, given that highways are currently prohibited to pedestrians, semi-autonomous (level 2 and level 3) vehicles could (normally) drive on highways without encountering pedestrians. Another comment in *Other* mentions consumer trust in autonomous vehicles. The author of comment number 43 says, *"you bought the car so you should be able [to] count on it saving your life."* Other comments in that category highlighted the mental consequences, such as post-traumatic stress disorder (PTSD) that passengers would suffer after crashing into pedestrians. For example, in comment 46 the author says, *"You also have to take trauma into account. If your car ran over some people, and you had no way to stop it? That would bring up some serious PTSD stuff."* Additionally, some commenters discussed the ability of the owner or passengers to have manual control in autonomous vehicles to solve such problems. For instance, in comment number 48 the author says, *"Perhaps, in the future, these self-driving cars would have some sort of 'manual override' if you want to make a split-second decision."* The last example is a suggestion that the buyer of the autonomous vehicle could

sign a contract that shows his/her preference of sacrificing 'utilitarian or deontological ethics'. For instance, author of comment 60 says, *"probably should have a wavier when you buy the car where the customer has to choose which rule applies first 'protect self' or 'least casualties'."*

5.3.3.1. Correlations in The Game Comments

As I mentioned previously, the multiple codes applied to some comments means that there may be more than one category and more than one reason for any given comment (as shown in Table 13). Some commenters had more thoughts than others; therefore, they posted more opinions. Some of those thoughts are related, even if they are from different categories. For example, the comment in row 3 of Table 13 has three opinions that fall under three categories: *"The car is at fault; the people in the car should bear the responsibility for the fault. In every case, I'd rather the car hit the wall than hit a pedestrian. This is a stupid test in any case. So many ways in practice to avoid this dilemma."* The commenter said that it was the car's fault and that the passengers should be held responsible. The category is *Sacrifice Passengers* and the reason is *Their Choice to Ride*. However, this commenter preferred the car to hit a wall, not the pedestrians. The category is therefore *Alternative Solutions* and the reason is *Swerve*. In the end, the commenter did not like the idea of The Game and the category is *Refraining* and the reason is *The Dilemma is Unrealistic*.

Out of 127 comments there are 62 comments that have more than one opinion, and these 62 comments have 160 opinions, out of a total of 225. Therefore, the reasons that occur together are correlated. Thus, for each comment I wanted to see how often reasons are mentioned together. To perform this task in Excel, I needed to calculate the combinations for each comment that has more than one reason, which is provided by the binomial coefficient:

$$\frac{n!}{r!\,(n-r)!}$$

Where n is the number of reasons and r is the number of reasons in each set of combinations under consederation. For example, the comment in row 3 in Table 13 has three opinions (n) which are 1.2, 3.1, and 4.1. These three codes would produce three combinations based on the formula. Each one of the combinations contains two codes (r), which are {1.2,3.1}, {1.2,4.1}, and {3.1,4.1}. Then, I used Excel to calculate the frequency for each combination, as illustrated in Table 15. In the previous example and as Table 15 illustrates, reasons 1.2 and 3.1 are mentioned together four times, reasons 1.2 and 4.1 are mentioned together two times, and reasons 3.1 and 4.1 are mentioned together two times. These combinations give a clear indication of the most common reasons that commenters posted together.

Table 15 The Game's Correlations

	1.1	1.2	1.3	2.1	2.2	3.1	3.2	3.3	3.4	4.1	4.2	4.3	4.4	4.5
1.1		0	0	1	0	0	1	0	0	0	1	1	0	0
1.2			0	1	0	4	2	3	1	2	4	1	0	1
1.3				0	0	0	0	0	0	1	0	0	0	0
2.1					0	0	1	4	1	1	5	1	2	2
2.2						0	6	1	0	0	2	0	0	0
3.1							0	4	1	2	2	1	0	1
3.2								8	3	0	4	1	1	2
3.3									6	1	10	2	3	7
3.4										2	4	1	1	1
4.1											5	2	3	0
4.2												7	3	3
4.3													1	1
4.4														0
4.5														

In general, some of these reasons are correlated. Table 15 shows the correlations and how strongly they are connected. The highest values in the table, which are (3.2-2.2), (3.3-3.2), (3.4-3.3), (4.2-3.3), and (4.3-4.2), show the most highly correlated reasons. The first two correlated reasons, *Keep Straight* and *Can Avoid* (3.2-2.2), were repeated six times. The commenters mentioned these reasons together because if the vehicle was programmed to stay in its lane during an unavoidable crash, then this action would be predictable. Pedestrians would thus have the opportunity to avoid the accident, especially when the vehicle warned them by sounding an alarm or honking. The second pair of correlated reasons are *Reduce Deaths* and *Keep Straight* (3.3-3.2). These reasons were repeated eight times. This correlation has a similar underlying motivation as the previous one which is to reduce deaths. Keeping straight would give others the chance to avoid the crash and this would be a way to reduce deaths. The third pair of correlated reasons of *Reduce Deaths* and *Disable the Vehicle* (3.4-3.3) was found six times. Disabling the vehicle includes such actions as shutting down the engine or activating the emergency brake, which would reduce the number of deaths by protecting passengers and avoiding pedestrians. The next correlation is *Machines Shouldn't Decide* and *Reduce Deaths* (4.2-3.3). These two reasons were mentioned together ten times, which is the highest number among all correlations. The commenters said that the vehicle must follow the law and should not make life-and-death decisions, because simply following the law will lead to a reduction of deaths in an unavoidable crash. The last correlated reasons of *Machines Shouldn't* Decide and *No Trust* (4.3-4.2) were repeated seven times. This correlation was mentioned previously when these reasons were combined to obtain the most repeated reason of unacceptability.

5.4. The Article's Comment Analysis

In this section, based on the coding process that I followed to code The Game comments, I explain how I created The Article's comment categories and codes. By following the process, I used with The Game comments, I divided all article comments into three main categories: *Unrealistic*, *Solutions*, and *Fear*. These three categories originated from the nature of the source — The Article — and the variety of comments. To clarify, several commenters did not accept the idea of letting a vehicle decide who would die, and so I placed those types of comments into the category of *Unrealistic*. The second group of commenters had accepted the idea of a vehicle making ethical decisions, and so they suggested some *Solutions*. The last group of commenters had *Fear* of the idea that a machine would make life and death decisions. Each group of commenters had different reasons that motivated them to choose their category. Therefore, I divided each category into several reasons, as shown in Table 16.

Table 16 Categories of The Article Comments

Category	Reason (Code)
Unrealistic	Machines are Better
	Uncertainty
	Cars are Safe
	Protect Passengers
	Greatest Happiness
Solutions	Designated Areas
	Safety Features
	Owner's Choice
	Warn
Fear	No Trust
	Machines Shouldn't Decide
	Unfair
	Crimes
	Other

The commenters who did not accept the dilemma and found it *Unrealistic* have five reasons. Some commenters said that *Machines are Better* than humans in general and so machines

would not face such a problem. The second group said that accident outcomes are unpredictable, and so there is *Uncertainty*. Essentially, there are too many variables to predict who would or would not die in a crash. The third group said that *Cars are Safe* with only a very slight chance of failure, claiming that cars have many safety features which will protect passengers in crashes. The fourth group said that the dilemma is unrealistic because the main duty of the car is *Protecting Passengers*, and this protection of passengers is the priority no matter what happens. The last group said that designing utilitarian vehicles that take the greater social good of *Greatest Happiness*, such as minimizing the overall number of accidents over time, is a good idea.

The group of commenters who suggested alternative *Solutions* have four reasons. The first group said that autonomous vehicles should only be allowed in *Designated Areas* that have no pedestrians, such as highways. The next group suggested that autonomous vehicles should be equipped with more *Safety Features* to be prepared for such dilemmas. Another group suggested giving *Owners* the *Choice* of who to sacrifice, either during the purchase process (in the contract) or after purchasing the vehicle while the vehicle is in the driving mode. The last commenters suggested that if the vehicle fails and faces an unavoidable crash, it should *Warn* others in the road to avoid it.

The last group of commenters who *Fear* autonomous vehicles have five reasons. The first group would *Not Trust* such vehicles and would not buy them. The second group of commenters said that *Machines Should Not Decide.* Another group of commenters said that it is *Unfair* to sacrifice innocent people because of others' mistakes, such as swerving to hit a wall and sacrificing passengers because of jaywalkers. The next group of commenters said that if autonomous vehicles were programmed to kill fewer people — sacrificing passengers

because they are fewer than pedestrians — then criminals and suicidal people will use those vehicles on roads as tools to commit *Crimes*. For example, a criminal could put a bunch of mannequins in front of the vehicle to make it swerve and induce it to kill the vehicle's passengers instead of the mannequins. The last group of commenters have *Other* opinions that fall into the *Fear* category. For example, some of these commenters said that these types of dilemmas will slow down the development process of autonomous vehicles; furthermore, these types of articles create fear among the public.

5.4.1. Recoding The Article Comments

For the purpose of calculating the magnitude and patterns of the categories and reasons, I recoded The Article comments using the same strategy that I used to recode The Game comments. This strategy, which is illustrated in the section entitled Recoding The Game Comments, was used to produce Table 17. For example, beside the comment that verbalizes unacceptance of the trolley problem in autonomous vehicles because it is *unrealistic* since crash outcomes are unpredictable and *Uncertain*, I put code number 1 for the category and code number 1.2 for the reason. I recoded all comments by this method and obtained a complete set of comments, summaries, category natural numbers, and reason rational numbers as the sample illustrates on Table 18 (all comments found in Appendix B).

Table 17 The Article's Numerical Symbols

Category	Numerical Symbols	Reason	Numerical Symbols
Unrealistic	1	Machines are Better	1.1
		Uncertainty	1.2
		Cars are Safe	1.3
		Protect Passengers	1.4
		Greatest Happiness	1.5
Solutions	2	Designated Areas	2.1
		Safety Features	2.2
		Owner's Choice	2.3
		Warn	2.4
Fear	3	No Trust	3.1
		Machines Shouldn't Decide	3.2
		Unfair	3.3
		Crimes	3.4
		Other	3.5

Table 18 The Article's Recoding Sample

N	Comment	Summary	Category	Reason
1	A self-driving car's first duty is to protect its passengers.	Protect Passengers	1	1.4
2	I think self driving should only be allowed on major highways and freeways. City driving is far to[o] variable for autonomous vehicles.	Self-driving cars should be on highways only.	2	2.1
3	So what if some of those little twerps who drop big rocks off of highway overpasses decide to get together with a few of their friends and run out in the road, "for the lulz?" What are those ten people doing in the road anyway? Give the driver control.	What if someone throws something in front of the car? Let the driver decide.	3 2	3.4 2.3
4	…	…	…	…

5.4.2. Calculating The Article Comments

I calculated the code frequencies and percentages for The Article comments using the same strategy as the one I used in Calculating The Game Comments to produce Table 19.

5.4.3. Results of The Article Analysis

Table 19 The Article's Calculation Results

Category	Frequency	%	Reason	Code	Frequency	%	% out of 241
Unrealistic	95	39.4%	Machines are Better	1.1	29	30.5%	12.0%
			Uncertainty	1.2	25	26.3%	10.4%
			Cars are Safe	1.3	9	9.5%	3.7%
			Protect Passengers	1.4	16	16.8%	6.6%
			Greatest Happiness	1.5	16	16.8%	6.6%
Solutions	57	23.7%	Designated Areas	2.1	9	15.8%	3.7%
			Safety Features	2.2	21	36.8%	8.7%
			Owner's Choice	2.3	16	28.1%	6.6%
			Warn	2.4	11	19.3%	4.6%
Fear	89	36.9%	No Trust	3.1	18	20.2%	7.5%
			Machines Shouldn't Decide	3.2	10	11.2%	4.1%
			Unfair	3.3	20	22.5%	8.3%
			Crimes	3.4	15	16.9%	6.2%
			Other	3.5	26	29.2%	10.8%
Total	241	100%			241		100%

Table 19 displays the results of The Article's comment analysis. For a verity of reasons, 39.4% of commenters found the dilemma to be *Unrealistic*. 23.7% suggested some *Solutions* to solve the problem or to prevent the problem from happening, while 36.9% had a *Fear* of autonomous vehicles that make ethical decisions.

Let us go into each category in more detail to examine why commenters have commented differently. Of the commenters who found the dilemma *Unrealistic*, 30.5% said that *Machines are Better* than current vehicles and humans at driving. Autonomous vehicles will never face such a problem because they will be programmed to follow the law, equipped with sensors and radars that make them react faster, and connected to the environment and infrastructure. For instance, the following author of comment number 65 (see the Appendix B) refused the dilemma because such vehicles have advanced technologies: *"This isn't even a real problem. The car/computer/set of sensors is/are better equipped to avoid these situations then people*

are. By a lot." Additionally, some commenters said that these types of vehicles are better than current vehicles at self-checking and diagnosing; therefore, they will refuse to run in the first place if there is a potential brake failure. For example, in comment number 78 the commenter says,

> *...it can be programmed to refuse to operate if its owner has delayed important maintenance to the point where the car poses a risk to riders or pedestrians. A car driven by sophisticated computers may also be able to run complicated diagnostics on itself that current cars cannot, or detect problems before there is a complete brake failure.*

Another reason that machines will behave better than humans in current vehicles is their highly advanced systems that communicate with the road and traffic infrastructures and with each other. These advanced systems allow them to detect danger before humans can. In comment number 68 the commenter says, *"This is what road side units and network infrastructure are for, these will detect crossing pedestrians earlier than the car can, communicate that to the car which will slow down before such a scenario occurs."*

The second reason that motivated commenters to find the dilemma *Unrealistic* is that the outcomes of the potential crash are *Uncertain*. This reason represents 26.3% of *Unrealistic* comments. For example, author of comment number 4 says,

> *...it presumes that the result of crashing into the wall can be known, that such a crash will certainly kill the occupants. That will depend on a lot of variables including at least the speed of the car, the makeup of the wall, the angle off (sic) impact and the safety features of the car.*

As another example, author of comment number 48 says,

> *This whole article is based on a false dichotomy: Kill others or kill yourself. The choice really is hit somebody or try to avoid them. You DON'T know if they are going to die if you hit them. You DON'T know if you're going to die if you hit a wall.*

The third reason in the *Unrealistic* category is *Cars are Safe* (9.5%), which affirms that passengers will be safe in crashes. For example, in comment number 133, the author says, *"Self-driving cars may be designed to higher passenger safety standards so that in the event they have to risk the well-being of the occupants they are more likely to survive without serious injury anyway."*

The next reason in this category is *Protect Passengers*, which represents 16.8% of *Unrealistic* comments. Some commenters found that the main duty of autonomous vehicles is to protect passengers. For example, in comment number 3 the author writes, *"A self-driving car's first duty is to protect its passengers."* In comment number 14 the author says, *"Protect the driver. Is no dilemma. If this situations arise, it is only because the people didn't respected the rules and ended up in wrong places. Why punish the innocent driver for mistakes of the others?"* The last reason in this category is *Greatest Happiness*, which represents 16.8% of *Unrealistic* comments. Commenters found that in general such technologies will save lives. For example, in comment number 46 the author writes, *"...with the exceedingly poor driving habits people have, and the war-like attitude of most drivers on the road today, having a technology like this will obviously save lives."*

The second category is a group of commenters who suggested alternative *Solutions* to stop the problem from happening or to at least reduce damages if the problem occurred. 15.8% of them

said autonomous vehicles should be allowed in *Designated Areas* only that have no pedestrians. For example, in comment number 11 the author writes, *"I think self driving should only be allowed on major highways and freeways. City driving is far to (sic) variable for autonomous vehicles."* Secondly, 36.8% suggested adding more *Safety Features* in vehicles to save the lives of passengers and pedestrians. Some of these suggestions, though infeasible, were mentioned many times such as adding airbags to the vehicle's exterior to protect pedestrians or adding ejector seats like those in jet fighters which eject passengers outside the vehicle before the accident occurs in order to save their lives. For instance, in comment 118 the commenter says, *"The ejection seat with a parachute solves this problem. Eject the occupants of the car, car crashes into wall. No one died."* For the next *Solution* in this category, 28.1% suggested that it be the *Owner's Choice* to choose their vehicle's mode preference of either self-sacrificing or full protection. One benefit of this solution lies in assigning responsibility. For example, in comment number 64 the author writes, *"let the car owner decide what kind of algorithm is preinstalled in their vehicle, the 'sacrificing the passengers type' or the 'killing the unfortunate pedestrians type' and assume legal/ethical responsibility accordingly."* The last suggestion which was posited in 19.3% of comments in *Solutions* is to allow the vehicle that will inevitably crash to *Warn* others, especially pedestrians, by sounding alarms to give them a chance to run away or avoid the vehicle. For instance, in comment number 140 the author asks, *"Did anyone consider using a horn?"*

The third and last category is *Fear* which has five reasons. 20.2% of commenters who fear autonomous vehicles would have *No Trust* if the vehicles are known to sacrifice their passengers. For example, the commenter of comment number 26 says, *"Almost nobody will buy a car that is programmed so sacrifice the owner. So, beware pedestrians!"* Secondly,

11.2% of commenters said that *Machines Shouldn't Decide*. Those commenters are scared of autonomous vehicles because of the potential failures that are caused by eliminating humans from the loop. For instance, in comment number 111 the author reflects, *"What if the computer fails (and it will)? People will become lazy behind the wheel. Humans should NEVER be replaced completely by computer when it comes to things that can take or save human life."*

Next, 22.5% of commenters found that it is *Unfair* to sacrifice innocent passengers inside the vehicle as a result of the behaviours of jaywalkers or animals in the roads. For instance, the author of comment number 76 concludes, *"Clearly this is ridiculous. What if it's 5 deer and the driver dies for deer?"* In the same context, criminals could take advantage. If criminals know that autonomous vehicles are programmed to avoid killing pedestrians, they could find ways to commit *Crimes* (16.9%) which force such vehicles to kill its occupants such as by swerving when a criminal jumps in front of the vehicle. For example, in comment number 39 the author writes, *"I can see hoodlums stepping out in front of cars to cause car crashes if they're programmed to avoid pedestrians at all costs."* As another example, in comment number 13 the commenter says, *"one could perhaps imagine a 'group of thugs' leaping into the road on purpose in order [to] cause this situation and kill the driver."* If autonomous vehicles are programmed to sacrifice fewer people, then criminals will take advantage of this feature to commit *Crimes* toward passengers inside the vehicle by jumping or blocking the road as groups. Clearly, this is *Unfair*. Therefore, we may add those two reasons' percentages together to conclude that *Unfair* plus *Crimes* (22.5% + 16.9%) will be 38.4% out of all *Fear* comments. In fact, adding these percentages together produces the highest number of reasons out of all the comments: *Unfair* represents 8.3% out of all comments in all categories and *Crimes* represents 6.2% out of all comments. Combined together these reasons comprise

14.5% of the 241 comments. This reason of *Unfair* and *Crimes* is higher than the first reason in the first category, *Machines are Better*, which weighs 12% out of all comments. To conclude, people are afraid that such vehicles will kill innocent people or be used as a tool to kill people.

The last reason in the *Fear* category is *Other*, which represents 29.2%. This category contains comments that also show a diverse collection of fear-related issues. For instance, some commenters did not like the very idea of The Article because such articles generate fear among people. In comment number 44 the author critiques, *"this is a terrible article and is attempting to create fear amongst the uninformed public."*

Other commenters were concerned about hacking. For instance, in comment number 70 the author says, *"Well, we all know that EVERY type of programmed device can be hacked.. so, what about individuals or organizations who decide to kill people intentionally?"* Another commenter worried about hijacking, for example in comment number 99 the author writes, *"What if it is a purposeful hijack situation where a person or persons with weapons, intent on taking the car or occupants of the car by force..or killing them? How will the car handle that?"*

5.4.3.1. Correlations in The Article Comments

As illustrated in Table 18, the Simultaneous Codes in some comments result in more than one category and reason within one comment. Out of 160 comments there are 61 comments, which have 142 opinions out of 241 opinions. To obtain the correlations, I calculated the combinations in each comment to see how often reasons are mentioned together. I repeated the same steps that were conducted in Correlations in The Game Comments. The calculation of combination frequencies is shown in Table 20.

Table 20 The Article's Correlations

	1.1	1.2	1.3	1.4	1.5	2.1	2.2	2.3	2.4	3.1	3.2	3.3	3.4	3.5
1.1		4	2	2	3	3	3	0	3	0	2	3	3	1
1.2			3	1	1	0	1	0	4	1	2	4	4	1
1.3				0	1	0	1	0	1	0	0	1	2	0
1.4					2	2	0	0	0	1	2	4	0	0
1.5						0	2	0	0	2	0	2	0	1
2.1							0	0	0	0	0	1	0	0
2.2								0	2	0	0	1	1	1
2.3									1	0	1	1	1	2
2.4										0	1	0	3	0
3.1											2	2	1	1
3.2												0	1	2
3.3													3	0
3.4														1
3.5														

Certainly, as I stated before, some reasons have correlations with one other. Table 20 shows the correlations and how strongly they are connected. For example, the main duty of the vehicle is *Protecting Passengers*, therefore it is *Unfair* to sacrifice them because of jaywalkers. Table 20 displays all correlations between every pair of reasons as well as how many times reasons are mentioned in one comment. While there is no disparity in the number of correlations on the table, four reasons have been frequently repeated with others: 1.1, 1.2, 3.3, and 3.4 (*Machines are Better, Uncertainty, Unfair*, and *Crimes*). Therefore, we could say that the largest two groups of participants are as follows:

(1) A group who finds that the dilemma is unrealistic. It is uncertain whether crash outcomes will lead to killing passengers or pedestrians, given that autonomous vehicles are better than current vehicles and human drivers at making fast decisions.

(2) A group who are afraid that if autonomous vehicles are programmed to avoid pedestrians at all costs, it would be unfair to sacrifice innocent passengers inside the vehicle since it is pedestrians who may be flouting the law. In addition, this capability of making ethical choices could lead to humans to commit new types of crimes.

5.5. Discussion

Analyzing both data sets answers two of my research questions that focus on public acceptance and trust. The questions are: 1) How do ethical decision-making algorithms need to be constrained in order to make them acceptable to the public? 2) If humans are to trust robots to make ethical decisions, what are the conditions under which this trust is justified?

Let us first summarize the results of the analysis in general and then discuss the answers to those questions. As illustrated in Figure 13, in the first set of comments for The Game few people supported the idea of sacrificing passengers or pedestrians, even though The Game depends on that sacrifice. The majority of participants either suggested solutions that solve such dilemmas or refrained and refused the idea of The Game. Commenters who suggested solutions believed that they have some ideas to contribute to preventing trolley-type dilemmas from happening. At the same time, they believed that there are more options than the two killing options presented by The Game. The extremely loaded question of The Game led participants to refrain from accepting the premise of the dilemma because it forces participants to choose between two evils as if there were no other choices. Potential crashes have many possible outcomes and nobody can predict the actual consequences. Therefore, the results of the actions are uncertain and nobody can be certain of their consequences.

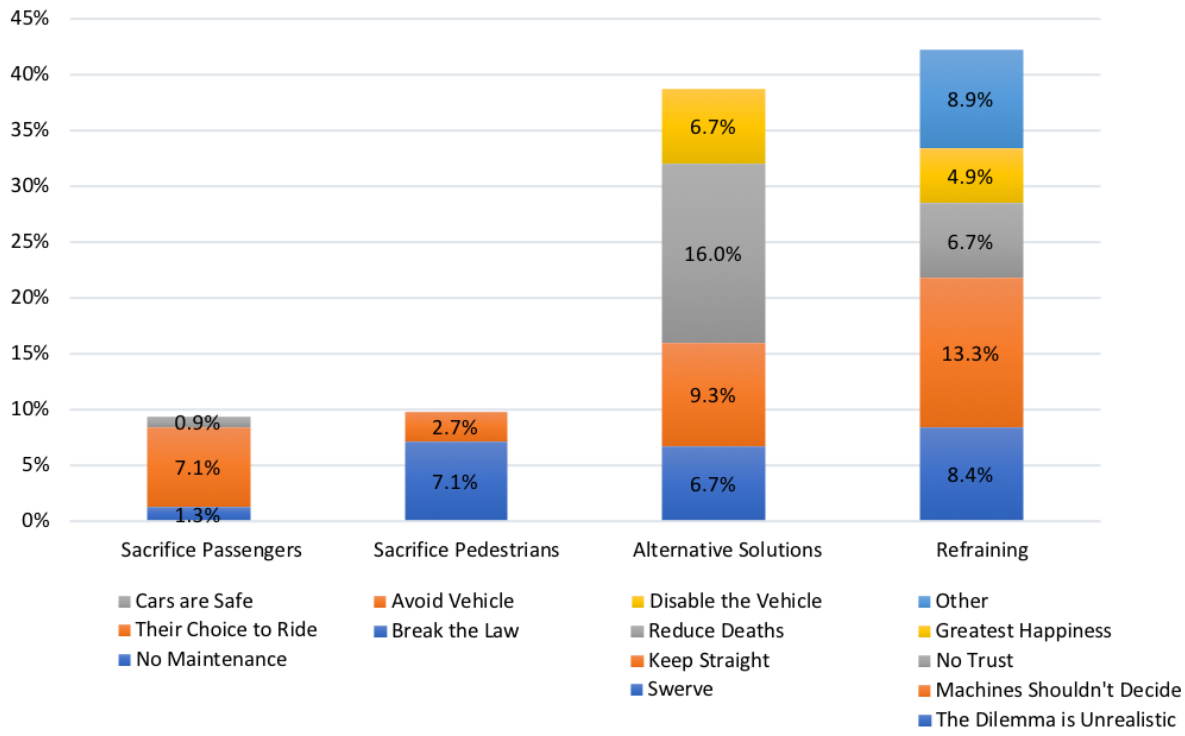

Figure 13 The Game's Comment Calculation Results

For The Article, there are very few sacrificing comments because of the nature of The Article which explains the ethical dilemma of algorithmic morality. Unlike the players of The Game, the readers of The Article are not asked to choose between two options. Therefore, participants either propose solutions or refuse the premise of the dilemma as illustrated in Figure 14. I believe that commenters who suggested some solutions have the same motivations as the first set of commenters. Yet the participants who refused to accept the premise of the dilemma did so either because they believed it is unrealistic or because they fear autonomous vehicles. It could be that this concern is amplified by the portrayal of intelligent machines in popular culture such as sci-fi movies. In addition, some articles published in the media have reinforced a fear narrative in public awareness as we discussed earlier in Results of The Article section. One consequence of this anxiety about autonomous machines is that it could slow down the acceptance rate of and trust in autonomous vehicles. Governments, automobile corporations,

and technology companies need to work consistently together to make autonomous robots, including autonomous vehicles, become acceptable and trustworthy. The transparency model discussed in the previous chapter is one solution to this problem.

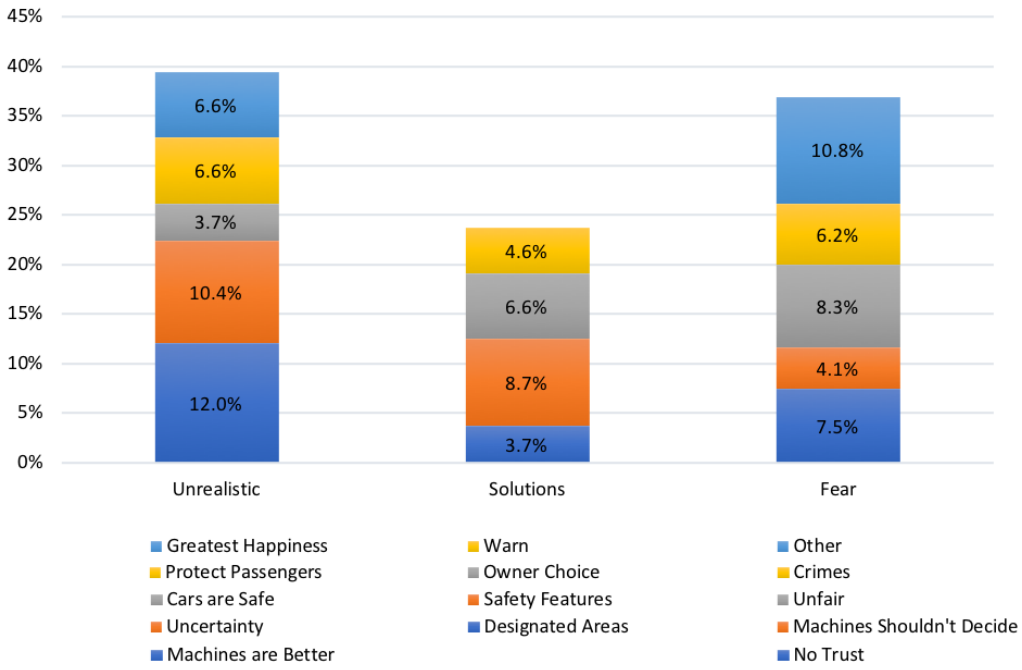

Figure 14 The Article's Comment Calculation Results

To gain public acceptance, autonomous vehicles need to be transparent in terms of their decision-making process and they need to act predictably on roads. The results and correlations tables for both sets of comments show that people do not want machines to make life-and-death decisions; however, if machines are programmed to make such decisions, then they must be programmed to reduce the total number of deaths (utilitarianism) by acting in predictable ways (transparency). Yet people fear unexpected uses of autonomous vehicles that may lead to crimes that take advantage of the nature of an autonomous vehicle's programming. Furthermore, people have doubts about whether to buy such vehicles because they not yet trust autonomous vehicles. For example, in Table 15, we can observe that some reasons in the

correlations table have high values and are correlated with other reasons, especially reasons 3.2, 3.3, and 4.2 (*Keep Straight, Reduce Deaths,* and *Machines Shouldn't Decide*). Therefore, participants want a predictable autonomous vehicle, which is one component of the transparency model. Additionally, they want a utilitarian vehicle; however, they do not accept the idea of machines making ethical decisions. In fact, Bonnefon et al. (2015) in their study found that while participants supported the decision of others to buy utilitarian vehicles that reduce deaths, they did not want to buy autonomous vehicles for themselves.

Many recently conducted surveys have asked a diversity of people about their feelings toward autonomous vehicles. For example, 75% of American drivers feel afraid to ride in a self-driving car (Stepp, 2016), and in another survey, 53% of respondents say that they are scared of autonomous vehicles (PWC, 2016). Seventy per cent of drivers in Britain not feel confident about riding in an autonomous vehicle (Loughran, 2016); and 63% of Canadians do not yet trust self-driving cars (Sanford, 2016). People are not yet ready to trust autonomous vehicles because they do not know how they work and they are afraid of new technology. On the other hand, 75% of American drivers who own cars with semi-autonomous technologies are more likely to trust the technology in their vehicles than people who do not own cars with such technologies (Stepp, 2016).

Purchasing a product does not mean that the product has been accepted. There are other players in the product's circle of influence such as members of society as a whole who live with the product and are affected by its outcomes. Acceptance is a relation between subject, object, and context (E. Fraedrich & Lenz, 2016). The subject, in this case can refer to drivers and owners but also to cyclists, pedestrians, society, engineers, developers, regulators, corporations, and governments. We can call these subjects the stakeholders that have a relationship to

autonomous vehicles. Object refers to the artifacts that will be used by the subject, which in this case is autonomous vehicles and context refers to the environment in which the subject uses the object.

The Automated and Connected Driving report by the Federal Ministry of Transport and Digital Infrastructure (BMVI) in Germany is in agreement with the idea of programming autonomous vehicles to reduce the total number of personal injuries in unavoidable crashes and in their doing so without any type of discrimination that allows vehicles to choose based on personal characteristics (BMVI, 2017). However, actual idea of reducing the number of injuries was not accepted unanimously, as reported by Luetge (2017): "This guideline was debated controversially, and it was not adopted unanimously by the committee's members" (p.552). In addition to the other guidelines, the report includes the guideline the vehicle should accept damage to animals and property if they conflict with humans. Moreover, according to the BMVI's report, autonomous vehicles should not sacrifice innocent people who are not involved in crashes. For example, a vehicle in an unavoidable crash should not be allowed to sacrifice a man walking on a sidewalk to save two pedestrians crossing the road who are the cause of the problem (BMVI, 2017).

In my opinion, autonomous vehicles should not make life-and-death decisions. If they are programmed to minimize the number of deaths in any given situation, people will not trust them, either as occupants or pedestrians. People will not use such vehicles if they know that they might be sacrificed in an accident. Otherwise, people will adapt their behaviours to avoid such vehicles, except in "herds", since by being in the presence of large numbers, they would be protected from harm. Thus, an algorithm designed to minimize deaths would lead to conflicts in society. For example, if an autonomous vehicle were to crash into either a safe

vehicle (e.g. Volvo) with one occupant or an old, small, and cheap vehicle with five occupants, what should the vehicle do? Based on an algorithm that aims to reduce deaths, it would target the safer vehicle for at least two reasons: (i) the vehicle is safe and so the occupant will likely survive the crash; and (ii) the vehicle has only one occupant, which is a fewer number of occupants than the other vehicle. This assumes, of course, that the vehicle is able to recognize car brands and identify the number of passengers in a vehicle, among other characteristics. Therefore, people who pay more to buy safer cars would be targeted, and this, at least to consumers of safe vehicles, would not be fair. This thought experiment has been raised before (e.g. Goodall, 2014) and would make such ethical dilemmas more complicated if autonomous vehicles were deployed in large numbers and were all programmed to reduce deaths in the way that the BMVI report suggested.

In unavoidable crash situations, attempting to reduce the number of deaths by changing lanes or swerving suddenly could increase the danger for the vehicle's occupants, other vehicles, and pedestrians. In such a situation, it would be preferable for an autonomous vehicle to simply follow the law and act in a predictable way, which means that it should not swerve, (i.e. keep driving in its lane) and should try to stop (as much as possible) before hitting any object. Additionally, the vehicle ought to engage in some sort of communication with other vehicles and pedestrians about its circumstances and intentions. The main factors of predictability and transparency play a very important role in dealing with such vehicles. If people know that autonomous vehicles are programmed to act predictably either in unavoidable crash situations or normal driving situations, then it will be easier for the public to accept and trust autonomous vehicles.

Yet acceptance and trust are not obtained only from transparency, as there are other factors as well. Transparency is also needed to demonstrate how autonomous vehicles work during operation. In addition, there are other factors affecting acceptance and trust that are required even before operational transparency. Before an autonomous vehicle is released, bought, or used, society, including end-users, need to be informed about their vehicle's specifications and functionality such as autonomy levels, embedded ethical principles (if any), driving modes (and the extent of their end-user control), as well as its communications features. These capabilities should be disclosed by vehicle corporations to potential users. Furthermore, in addition to transparency and disclosure, society as a whole, particularly in the legal and insurance sector, will demand that explanation be provided about how and why vehicles make certain decisions, particularly if an accident happens. Consequently, trust and acceptance are also affected by three other factors: disclosure, transparency, and explanation.

5.6. Content Analysis and Transparency Model Mapping

As I indicated in section 2.6, the function of content analysis in this study is to support my hypothesis about the importance of transparency. The transparency model components were arrived at from my analysis of the literature and I conducted content analysis to corroborate my ideas with empirical evidence. The transparency model has three main stages: disclosure, transparency, and explanation. In this section I discuss how I mapped some codes with those three stages. However, it is important to note that there are no codes (and hence no mention in discussion forums) related to the explanation stage. The components of my model have been determined analytically only from the literature review.

The codes that correspond to the disclosure and transparency stages are shown in Table 21. For example, the choice that vehicles owners/users have to ride the vehicle with full knowledge of its capabilities (and limitations) is a type of disclosure that must be known to them before they use or buy the vehicle such as the ethical principles that the vehicle rely on to sacrifice either passengers or pedestrians in unavoidable crash situations. In another example, potential buyers could be required to sign a contract that shows exactly what the vehicle's ethical principles are and what its abilities are to make decisions. This would be a declaration of the owner's choice to ride the vehicle and would constitute a certain kind of disclosure.

In terms of transparency, the "keeping straight" and "warning" codes refer to a predictable action that show people near the vehicle what it is doing so they can avoid the vehicle during that behaviour. At the same time people around the vehicle will be informed that the vehicle is analysing and sensing the environment around it. The "disabling the vehicle" and equipping it with more "safety features" is a kind of transparency which tells users that the vehicle is safe and that it is able to deal with the situation with minimal damage.

Table 21 Mapping Transparency Model and Codes

Purpose	Reason (Code)
Disclosure	Their Choice to Ride
	Cars are Safe
	Owner's Choice
	Machines Shouldn't Decide
Transparency	Avoid Vehicle
	Keep Straight
	Disable the Vehicle
	Designated Areas
	Safety Features
	Warn

5.7. Content Analysis Limitations

The Content Analysis Methodology (section 2.6) mentions two limitations to the content analysis: 1) the codes that emerged from the analysis are limited by the available content in the discussions and 2) that building a theory about a phenomenon by only using online published content is a limitation because it ignores other unpublished content about the same phenomenon.

While these two limitations were known *a priori*, before the content analysis was conducted, this study faced some additional limitations. First, some of the comments from the discussions had to be disregarded because they were not relevant to the scope of this thesis, such as those comments that mentioned geographical factors. Second, I chose to focus only on the initial comments in the discussion because they are the expression of the first impressions that the respondents had about the dilemma. Subsequent discussion comments (further in the thread) were always are affected by the ideas that others had expressed after the initial comments. Thus, the emergent codes for these comments cover only a portion of all the analysed text. A complete coding of all the text would have required more than twice as many codes, none of which would have been directly relevant to the respondent's initial reactions.

There were other limitations to this study as well. One such limitation is that it was not possible to validate the authenticity of the respondents' identities without initiating a direct contact with the respondents, such as conducting a survey. One advantage of performing this extra step would have been to determine the respondents' level of education and knowledge of decision-making machines. Using the current method, we are not able to ascertain how their views were formed.

6. Conclusion

Robots in the future will have a greater capacity to perform more tasks. Therefore, people will rely on robots and trust them to do many more jobs and many different kinds of jobs as well. A number of these jobs or tasks will be related to people's safety, privacy, health, and even lives. Consequently, robots must follow rules that enable them to be reliable, trustworthy, and good companions that maintain human safety at all times, such autonomous vehicles which will be obliged to follow ethical rules that protect people's safety. On the other hand, some types of robots do not need to follow ethical rules (such as vacuum robots). Therefore, not all robots should be required to obey ethical rules in the future. We do not need fully ethical robots to perform every single task that we assign to robots.

One of the main ideas in this thesis is that autonomous decision-making machines – ranging from autonomous vehicles to chatbots – are already able to make decisions that have ethical consequences. Hence, if such machines are eventually deployed on a large scale, then members of society must be able to trust the decisions that are made by these machines. Establishing society's trust will require that these decisions are controlled by socially accepted ethical principles and that these principles and their role in machine decision-making are transparent. Indeed, one critical aspect of the trust relationship between humans and decision-making machines is the transparency of the mechanisms involved in decision-making and the explanations that can be provided for these decisions. This need for transparency and explanation is compounded by the corporate social responsibility of robot manufacturing companies to design these robots in ways that make them trustworthy. Members of society

will not trust a robot that works in mysterious, ambiguous, or inexplicable ways, particularly if robots must make decisions that are based on ethical principles.

The current literature on embedding ethics in robots is sparse. This thesis aims to partially fill this gap in order to help different stakeholders (including policy makers, industry, designers, and the general public) to understand the many dimensions of machine-executable ethics. To this end, I provide a framework (Figure 7 in Chapter 4) for understanding the relationships among different stakeholders who legislate, create, deploy, and use robots and their respective reasons for requiring transparency and explanations. The framework I propose aims to provide an account of the relationships between the transparency of the decision-making process in ethical robots, explanations for their behaviour, and the individual and social trust that results. This thesis also presents a model that decomposes the stages of ethical decision-making into their elementary components (Figure 8 in Chapter 4) with a view toward enabling stakeholders to allocate the responsibility for such choices. In addition, I propose a model for transparency (Figure 9 in Chapter 4) which demonstrates the importance of and relationship between disclosure, transparency, and explanation which are needed by stakeholders in society to accept and trust robots.

One important stakeholder is the general public and, in addition to providing an analytical framework with which to conceptualize ethical decision-making, this thesis also performs an analysis of the opinions voiced in hundreds of comments posted on public forums concerning the behaviour of socially autonomous robots. This analysis offers insights into the layperson's responses to machines that make decisions and provides support for policy recommendations that should be considered by regulators in the future.

6.1. Thesis Contributions

In this thesis I have provided a narrative literature review which covers various aspects of robot ethics, beginning with defining robots and robot ethics in general. I have explored the ethical theories that can be applied to robot ethics, the notion of ethical responsibility (including corporate social responsibility), how these concepts apply to the problems of autonomous robots including autonomous vehicles in ethical decision-making situations, and the issues that arise from machine learning and corporate social responsibility. From the literature review I have identified certain areas which need further research, especially with regard to the relationships between the concepts of transparency, explanation, and trust. For this I have developed a conceptual model. The content analysis of human discussions about the ethical dilemmas faced by autonomous vehicles is a unique work. This work provides evidence which supports some of my ideas, such as the concept that transparency enables trust. Consequently, I have concluded that there is a need for transparency in the robotics industry because transparency affects society's trust in robots. I believe that this need can be understood effectively within a trust framework, a decision-making model, a transparency model, and a robot-capability labelling scheme.

6.2. Answering the Research Questions

As illustrated in Table 22, the ethical decision-making model has answered the first research question about situations that are not suitable for robots to make ethical decisions. The transparency model presented in Figure 9 has answered the second and fourth questions, respectively, which asked about the conditions that must be met to enable acceptance and trust

of decision-making robots. The trust framework presented in Figure 7 has answered the third question by addressing the role of explanation to enable trust and allocate responsibility.

Table 22 Answers to Research Questions

N	Research Question	Answer Provided by	Answers
1	In what situations and why might it be inadvisable to enable robots to make ethical decisions?	Ethical Decision-Making Model	- Making social autonomous robots without ethical guidance - Making self-learning robots without transparency
2	If humans are to trust robots to make ethical decisions, what are the conditions under which this trust is justified?	Transparency Model	- Presence of disclosure, transparency, and explanation
4	How do ethical decision-making algorithms need to be constrained in order to make them acceptable to the public?		- By ethical rules, social values, and laws
3	If decisions by robots are made using ethical principles, what is the role of explanations in establishing trust and allocating responsibility?	Trust Framework	- Presence of explanation is a condition to enable trust

The decision-making model (Figure 8) for ethical robots discussed in Chapter 4 provides an answer to the first question. Since there is currently an absence of ethical guidance in robots that make decisions and there is a lack of ethical considerations in the design of autonomous robot systems, robots' decisions must be guided and controlled by machine-executable ethical principles and values. I have found that it is crucial to expand the general robot paradigm – the Sense-Think-Act paradigm – and refine the decision-making model in robots to include the process of ethical guidance. Robots that operate using utilitarian calculations or deontological rules both need ethics to be embedded in their decision-making process. These ethical guidance systems are essential for controlling robots' decisions so that they comply with laws, rules, ethics, and social values.

Social autonomous robot that are be deployed without ethical guidance could make mistakes that have unwanted ethical consequences. Therefore, it is advisable to create social

autonomous robots with embedded ethical principles and social values. Yet it is important to note that different robots may have various levels of impact if they made ethical mistakes. For example, a chatbot that posts unwanted messages that affects the emotional well-being of only a few people is not comparable to an autonomous vehicle which makes decisions that weigh the lives and health of passengers and pedestrians in unavoidable crash situations. Both actions could be immoral and lead to unethical consequences; however, killing people is far more serious than offending people verbally.

Next, the answer to the third research question has been illustrated by the trust framework (Figure 7) which shows the relationships between the different components of explanation and trust. The main factors in the robot industry consist of robot manufacturers, regulators, robots themselves, and society. The framework that was elaborated in Chapter 5 and identifies the role of each factor and its relationship with others for the allocation of responsibility for robot decisions that in turn leads to relationships of trust. This framework highlights the robot companies' responsibilities toward their stakeholders, especially users. Users' trust in their robots is enabled by the transparency of the robots' operations, which in turn enables explanations for the robots' behaviours. Figure 7 defines the relationship between those three components as *corporations* create *robots* for *society*. To clarify and explain the relationship between those three components, I argue that in the case of an error or a mistake made by a robot, society as a whole, including users, needs the manufacturers to explain how and why that mistake happened. Explaining how and why requires an understanding of the robot's decision-making process. Therefore, I added a decision-making model component - Figure 8 - to the framework and linked this to the other components. My decision-making model helps to situate the necessity for robots to enhance the explanations that are needed by members of

society to trust such autonomous robots' decisions. However, explanation is not the only factor that enables trust in society: people also need to know what internal mechanisms are embedded in the robots to prevent them from making mistakes. Given that a machine whose decisions are controlled by ethical principles is designed to protect users from robots that might make ethical errors, members of society would need to know what types of ethical rules or principles are embedded in such robots. Accordingly, a robot must be transparent in terms of its ethical rules which means that these rules must be known to users and all members of society. Therefore, transparency and explanation are two necessary factors for society to trust robots' decisions. Furthermore, regulators must guide robotics companies to make ethical robots that are transparent and explainable to protect society from potential failures. This combination enables trust in robots among society members, with trust being a very important factor that affects companies' sustainability in markets. In addition to being responsible for providing transparency and explanation, companies must use transparent and explainable computing methods to embed ethical rules and to implement ethical decision-making processes in robots.

The second and fourth research questions are answered by the transparency model (Figure 9) which demonstrates society's requirements for accepting and trusting robots' decisions. In the literature the need for transparency is clear, yet there is an absence of detail about which aspects of a robot's operation need to be transparent and explainable. I wanted to ensure that the interaction between transparency and explanation is clear. As well, the literature does not sufficiently enable an understanding of the conditions under which members of society are prepared to accept and trust robots. To fill this lacuna, I performed a study of opinions expressed in discussion groups about social autonomous robots. I wanted to identify what people need in order to accept robots and trust their decisions. Based on my trust framework,

trust is a clear result of enabling transparency and explanations; however, I still did not know exactly what is the perceived need by members of the public. Therefore, I conducted a content analysis study to extract laypersons' opinions about social autonomous robots. This study of people's opinions provided evidence to support the view that transparency leads to trust. People will not accept or trust robots that operate in mysterious ways. Therefore, I combined the results from the analysis and literature review to build a transparency model. I divided transparency into three stages, namely pre-operation, operation, and post-operation. Before users buy a robot, they need to have initial information about the robot and know how it operates. The proposal for this stage is the disclosure of information about how the robot can be expected to behave. In this stage, I introduced the idea of a disclosure label which would have to be presented to potential buyers in the same way as robots' technical specifications. During its operation, users and people who are affected by the robot's decisions need to know what the robot is doing in the field. Thus, the proposal for this stage is an operationally transparent robot. If the robot makes a mistake, users need to know why and how the mistake happened. Following operation, the purpose of the next stage is explanation. There is therefore a flow and an interaction between those three components that I have defined in the Transparency Model.

6.3. Challenges

I faced an important obstacle in writing this thesis: people do not have full knowledge about fully autonomous robots because these robots are not yet deployed in the marketplace; thus, it is difficult to rely on people's opinions about something that does not yet exist. I solved this problem in part by focusing on a specific type of autonomous robot – autonomous vehicles –

that is understood to some extent by laypersons because they have either read about it or currently use a semi-autonomous version of it.

Generally, people working in the field of robot ethics face a variety of challenges, such as whether it is even possible to compute ethics; how to resolve conflicts between ethical rules, cultural challenges, unpredictability and algorithmic bias. (Torrance, 2007) states the first challenge as "how to model ethics as a domain of rules or knowledge within intelligent systems" (p. 497). (M. Anderson & Anderson, 2007) ask this question: Are ethics the sort of thing that can be computed? Since there are many theories of ethics that have numerous facets, programmers need a variety of methods for their computation. For example, based on deontological theory, both the wrongness and the rightness of an action depend on the action itself, regardless of the consequences. But this mindset is a challenge if programmers want to apply this theory to a robot, because it avoids the consequences of the action. Sometimes a deontologically good action leads to a bad consequence, and vice versa. For example, in some situations a driver might break the law to save a human life, such as going through a red light to enable an ambulance to pass. But this raises the question of whether a programmer can implement this ethical decision-making method so that a robot behaves consistently in all situations. How could a programmer design an ethical robot to break laws and rules, but admit exceptions? A computational solution is required for the problem of default reasoning. Although such a solution has been researched extensively in artificial intelligence, it does not yet have a practical implementation in robots, especially not in utilitarian ones.

To offer another example, if a consequentialist theory such as utilitarianism is embedded in an autonomous robot, then the wrongness and rightness of an action would depend on the robot being able to predict and calculate the consequences of its actions. In this case, the

programmer's challenge is to compute and weigh the positive and the negative consequences. Could all the consequences even be enumerated, let alone evaluated? Also, merely obtaining information will not result in an ethical robot behaving ethically (M. Anderson & Anderson, 2007). As Tzafestas (2016b) shows, the theory of utilitarianism also presents many points that pose many challenges for the ethical robot. For instance, it is difficult to determine who might be affected by an action. In addition, one action could produce many results, which means that any given outcome is not necessarily the result of only one action. Therefore, uncertainty and unpredictability exist in the consequences of a robot's action. Moreover, it is very difficult to measure things like pleasure and other emotional reactions that would need to be known to maximize utility. The concern lies in the ability of a robot to perform tasks that need human intelligence, including learning, logic, and reasoning (Kernaghan, 2014). The overall process of determining the consequences of a possible action is context-dependent, complex, and time-consuming; some actions may have desirable consequences in one scenario but undesirable consequences in another. While some previously mentioned experiments have succeeded in calculating consequences, how to compute emotions, which some thinkers say is also necessary for computational ethics, is still unknown.

In any implementation of an autonomous ethical robot, there are additional challenges to those outlined above. For example, the ethical robot that runs on a bottom-up strategy needs more time to develop its operational ethical behaviours. Since it learns from its environment, it requires more time and likely needs to make more mistakes in its reactions with users in real life than a top-down programmed robot. Time and tolerance for initial mistakes are therefore constraints for bottom-up ethical robots. Another challenge in building an ethical robot or AMA with the bottom-up strategy is how to define a moral action for the system (Wallach,

2010). Wallach adds that using a bottom-up strategy, a robot could not reach the goal of maximizing the goodness because its goal could be vague; whereas, in the top-down strategy, the goal is defined and reachable.

The second challenge is possibility of conflicts between ethical rules. During the discussion of the first challenge I introduced an example which raised a question about how to compute an ethical robot that breaks laws, but only in some cases. This conflict sometimes occurs between deontological and utilitarian rules when the action/decision is unethical or illegal but has an ethical consequence. This conflict would likely be faced by utilitarian robots more often than deontological robots. Utilitarian robots rely on calculations of decisions' consequences and then choose the decision that has maximum benefits, whereas deontological robots simply decide and then act based on fixed rules regardless of a decision's consequences. However, deontological robots could find themselves operating in a different social context in which the rules that they are supposed to obey should no longer be applied. This could happen for instance when a social robot moves from one culture to another, or when a bot is implemented for audiences of a different culture.

The third challenge of ethical robot implementation is culture. Autonomy, dignity, and morality vary significantly from one culture to another, because every culture has different sets of social values (Tzafestas, 2016b). In addition, religious practices and traditions are not easily computed in machines (Wallach & Allen, 2009). Naturally, there are common values among all cultures, but at the same time there are many details and differences with highly complex applications in an ethical robot. For example, honesty and integrity are universal values shared among all cultures, and so social robots must be honest with their users. For example, a deontological rule is to always tell the truth and do not lie. However, how would a

social robot answer if a child asked it about the existence of Santa Claus? Moreover, within the same culture or religion, there are cultural discrepancies between certain groups' beliefs and actions, and sometimes conflicts between those groups can further increase the difficulty of formulating a uniform set of rules in an ethical robot. It is a challenging task for programmers of ethical robots to apply various rules or values that could conflict with one another. Furthermore, programmers must have extensive knowledge about the cultures of potential users of the ethical robot. It is impractical to design a social ethical robot for a specified community or culture without applying their particular values.

Similar challenges will arise when designing an ethical robot for different languages, because sometimes one word or gesture has extremely dissimilar meanings from one language/culture to another. Some words or gestures have a positive meaning in one language, but a negative connotation in another language. For example, while a thumps-up gesture is a common sign of approval in Western countries, this same gesture has the opposite meaning in some Eastern countries[28]. In addition, moving an ethical robot between different cultures could also be a problem because the robot will need reprogramming. That problem will be even greater if an ethical robot that runs based on a bottom-up strategy is moved to a different culture. This robot would have already learned ethics from the environment of one culture. After a geographical move, it would need to override its own built-in knowledge and build new knowledge based on its new experiences. Otherwise, it would risk performing tasks based on the first knowledge base, and such tasks could cause bias, misunderstanding, or conflict in a new context.

[28] Anderson, D., & Stuart, M. (2017, July 11). Hand gestures that could get you in serious trouble in other countries. Retrieved April 18, 2018, from http://www.businessinsider.com/hand-gestures-offensive-countries-world-2017-6

The last two challenges of unpredictability and algorithmic bias were discussed in the sections on Machine Learning and Algorithmic Bias in Chapter 3. Yet in general, robots that are equipped with machine learning ability have a high probability of behaving unpredictably. Users of such robots cannot predict what robots will do, even in the same situations and under the same circumstances. Even the robot creators cannot predict what the robot will do or explain what the robot has done. Therefore, if the robot is unpredictable or unexplainable, particularly when it comes to making ethical decisions, then these characteristics will erode the trust of users and the public. As I argued earlier, people will not trust robots that work in an ambiguous way. Note that explanations are not possible without transparency, but transparency is not sufficient to provide explanations. The user may be able to see everything about how a system functions and still not have an explanation for why it behaved a certain way in any given situation (e.g. AlphaGo).

Regarding the challenge of algorithmic bias, robots could make unwanted/unethical decisions because of the various biases in their decision-making systems. The biases originate from different sources such as biased training data, bias from the robot's interaction with users, bias from moving or deploying the robot in different environments, and emergent bias. Machine learning plays a significant role, especially in interaction bias and emergent bias. Robots could learn bad habits from their users and environment and then apply these bad habits back to their users, as what happened with Amazon's recommender system and Microsoft Twitter's chatbot Tay.

It is important to clarify that these challenges have not hindered me from recognizing the problems in robot ethics or from developing a general framework to solve some of these problems. Although people's knowledge about autonomous robots was limited in the content

analysis of Chapter 5, this limitation did not prevent me from understanding the layperson's need to trust autonomous robots.

6.4. Future Work

This thesis does not solve all of the problems nor fill all of the gaps in robot ethics. A great deal more work is needed in this relatively new area, especially experimental work on the interactions between people and autonomous decision-making robots. The following questions arising from this thesis could be the subject of further research:

1- Given that there are differences in social values and ethical rules across cultures, users may accept and trust robots that are regulated and controlled by their local ethical rules more readily than other robots that are controlled by international ethical rules. The presence of culture-dependent ethical rules in autonomous robots may affect the extent to which users are willing to trust these robots. Robot manufacturers could be encouraged to consider developing local ethical rules in the design of their robots to more quickly gain acceptance and trust in some markets.

2- The content analysis in Chapter 5 could be extended in complementary ways:

   a. In The Game case, for example, participants in a controlled study could first be asked to answer questions about machine-decisions in an instance of the trolley problem and then asked to answer some questions about what they think the role is of transparency, explanation, acceptance, and trust. This process could guide participants to formulate clearer opinions and disclose their real feelings about robots that make ethical decisions.

b. An experiment could be performed that tests people's reactions in real-world situations by asking participants to ride in an autonomous vehicle. Participants could be placed in situations where they face bullying from other drivers, ride in a vehicle that works in mysterious ways (i.e. with no transparency), have to choose between the lives of pedestrians or passengers, or must minimize the number of injuries and deaths in an unavoidable crash situation. Since this last experiment could have psychological side effects for the subject, an autonomous vehicle simulator would be better than a real-world experiment.

3- The trust framework could be extended:

a. A responsibility block could be added to the trust framework to cover more areas of robot ethics. Since responsibility has more than one component, these components could be linked to other components in the framework. For example, one of the research questions could be that, if we agree that we should assign responsibility to individual robots, what effect would this have on individuals in society and the corporate social responsibility of manufacturers?

b. In the trust framework, there is a block that shows ethics in robots. In the future, "ethics in society" could be added as a separate block with subcomponents that demonstrate the importance of individuals' ethical attitudes toward robots. This extension to the framework could influence how regulators could guide people, including users, to deal with robots, to protect robots, to protect users, and to protect society at the same time. The ethics of how people deal with and treat robots was first introduced in 2006 by Peter Asaro. An "ethics in society" block would illustrate that idea of human-robot interaction and its relationship with other components in my framework. Brooks (2017) demonstrates different

examples of behaviours by autonomous robot users which would generate social problems (see Societal Issues in Chapter 3). One benefit from an ethics in society block is the previously indicated need for laws that protect robots from the bullying behaviours of other users for the safety of robot users. Therefore, one of research questions that needs to be further studied is the question of how human-robot interactions are going to affect human socialization and the subsequent ethical behaviors? For instance, would treating autonomous decision-making robots like slaves habituate humans into treating other humans inhumanely?

c. The trust framework's computing block needs to be expanded. Software specialists could build a model that shows how to embed ethical principles and implement ethical decision-making rules in robots. Elements in an extended ethical computing model could be linked with other blocks and components in the framework.

4- The transparency model in Figure 9 could be expanded to have different levels. Transparency is not just one thing which must be taken as whole or not. Robotics companies could devise transparency criteria that support ethical requirements, predictability, and explanations without exposing their intellectual property or fully disclosing the details of their robots' internal processes. The interests of corporate intellectual property favour technological opacity; however, it may be possible to disclose enough information to satisfy transparency and explanation requirements without disclosing the intellectual property concerning the exact mechanism that controls the functioning of the robot. A future transparency model would show these different levels of transparency. Another pair of related research questions are: do

members of society need detailed explanations? Or is it sufficient to know that they exist, if needed, in a court of law, for example? Does the existence of these explanations affect the trust people have in these kinds of machines?

6.5. Epilogue

6.5.1. Deontological or Utilitarian Robot

Social robots are not evil, and they do not need to be evil to have harmful consequences on society. If they are built or used without built-in ethical constraints, then the social consequences could be significant. Social robots have been invented to make our lives easier, happier, and safer; however, to achieve these goals, they must be controlled by ethical rules. In addition, autonomous social robots should behave in a predictable and transparent way so as to not deceive users and to enable the proper allocation of responsibility for mistakes or accidents. Deontological robots might be better -in terms of predictability and transparency- than utilitarian robots, especially in autonomous vehicles, or at least in their initial deployment. Deontological robots are more easily transparent and more predictable, particularly if their ethical rules are disclosed by their manufacturers. On the other hand, utilitarian robots have more technological opacity, meaning that society, including users, are not able to predict their behaviours as easily and are hence less likely to trust them. This recommendation applies to robots that make decisions that affect people's lives, such as autonomous vehicles. This recommendation is less critical for entertainment robots that do not make life or death decisions.

6.5.2. Recommendations for Regulators

The trust framework, ethical decision-making model, and transparency model have been developed in this thesis in part to provide a foundation for recommendations to governments, including regulators and policy makers. In addition, corporations that allow robots to run in their platforms (e.g. social chatbots in Twitter and Facebook) must regulate their APIs (Application Programming Interfaces) so that they cannot be used in an unethical way. For instance, a recent study by Pew Research Center shows that 66% of the links to popular websites on Twitter are shared by automated bots[29]. Therefore, Twitter has started to restrict the automated bots that tweet, retweet, and like simultaneously from multiple accounts[30]. Although the automated bots are not necessarily bad, some of them have been used for harm, by, for example spreading fake news by dumping hundreds of identical tweets at the same time, or verbally attacking humans by following, tweeting, and retweeting them. Sometimes these bots are unrecognizable as bots by humans because humans cannot tell the difference between human users and automated bots. Thus, human users will follow bots, believe their news, and trust them. Therefore, when unmoderated by any ethical constraints these bots reduce credibility and effectiveness of online content.

6.5.3. Awareness is Needed

Increasing society's trust in social autonomous robots is a natural goal of all robotics corporations, but efforts to achieve that goal could lead to "overtrust" by users and result in problems in the future. By excessively relying on robots to perform too many tasks, people

[29] Wojcik, S., Messing, S., Smith, A., Rainie, L., & Hitlin, P. (2018, April 9). Bots in the Twittersphere. Retrieved April 13, 2018, from http://www.pewinternet.org/2018/04/09/bots-in-the-twittersphere/
[30] Roth, Y. (2018, February 21). Automation and the use of multiple accounts. Retrieved February 22, 2018, from https://blog.twitter.com/developer/en_us/topics/tips/2018/automation-and-the-use-of-multiple-accounts.html

may not consider the consequences for their privacy and safety. For example, users could use robots/bots to share their information or to obtain information without constraints. Similarly, drivers may rely on their semi-autonomous vehicles to perform most or all driving tasks and develop an excessive dependence on them. Overestimation – or overtrust – in autonomous robots could also lead to counterproductive societies. Moderation and awareness in using autonomous robots are key to maximizing their benefits. The role of regulators is essential for protecting societies, but more important is the role of individual responsibility in use of technologies. We are responsible as individuals for being aware of how we share our private information through robots/bots, how we trust shared information, and how we rely on autonomous robots.

In conclusion, the contributions of this thesis fill some gaps in the ethics and governance of artificial intelligence. The limitations of this work have not prevented me from understanding the relationships between different factors, namely regulators, robot corporations, robots, and societies. Understanding the role of each factor or component from the literature and conducting the content analysis helped me to define subcomponents for each factor and link these with other factors. This way of conceiving these relationships has supported me in building the trust framework, decision-making model, and transparency model. Ultimately, regulators are responsible for controlling the robotics industry to protect societies; robot companies are responsible for building ethical robots that are predictable, transparent, and explainable; and people are responsible for maintaining awareness of their use of robots.

References

Alaieri, F., & Vellino, A. (2016). Ethical Decision Making in Robots: Autonomy, Trust and Responsibility (pp. 159–168). Presented at the International Conference on Social Robotics, Cham: Springer International Publishing.

Alaieri, F., & Vellino, A. (2017). A Decision Making Model for Ethical (Ro)bots (pp. 203–207). Presented at the IEEE International Symposium on Robotics and Intelligent Sensors, Ottawa.

Allen, C., Smit, I., & Wallach, W. (2005). Artificial Morality: Top-down, Bottom-up, and Hybrid Approaches. *Ethics and Information Technology*, *7*(3), 149–155.

Allen, C., Wallach, W., & Smit, I. (2006). Why Machine Ethics? *Intelligent Systems, IEEE*, *21*(4), 12–17.

Alpaydin, E. (2010). Introduction to Machine Learning. (T. Dietterich, C. Bishop, D. Heckerman, M. Jordan, & M. Kearns, Eds.) (2nd ed.). London: The MIT Press.

Anderson, M., & Anderson, S. L. (2007). Machine Ethics: Creating an Ethical Intelligent Agent. *AI Magazine*, *28*(4), 15–26.

Anderson, S. L. (2011). Machine Metaethics. In M. Anderson & S. L. Anderson (Eds.), *Machine Ethics* (pp. 21–27). Cambridge: Cambridge University Press.

Angwin, J., Larso, J., Mattu, S., & Kirchner, L. (2016, May 23). Machine Bias. Retrieved November 1, 2016, from https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Arkin, R. C., Ulam, P., & Wagner, A. R. (2012). Moral Decision Making in Autonomous Systems: Enforcement, Moral Emotions, Dignity, Trust, and Deception. *Proceedings of the IEEE*, *100*(3), 571–589.

Asaro, P. M. (2006). What Should We Want From a Robot Ethic? *International Review of Information Ethics*, *6*, 9–16.

Asaro, P. M. (2008). How just could a robot war be. In A. Briggle, P. A. E. Brey, & K. Waelbers (Eds.), *Current Issues in computing and philosophy* (pp. 50–64). Amsterdam: Current issues in computing and philosophy.

Asaro, P. M. (2014). A Body to Kick, but Still No Soul to Damn: Legal Perspectives on Robotics. In P. Lin, K. Abney, & R. Jenkins (Eds.), *Robot Ethics: The Ethical and Social Implications of Robotics* (pp. 169–186). MIT Press.

Asimov, I. (1942). I, Robot.

Asimov, I. (1985). Robots and Empire.

Awad, E. (2017, May 18). *Moral Machines*. (I. Rahwan, Ed.). MIT.

Barnett, M. L., Jermier, J. M., & Lafferty, B. A. (2006). Corporate Reputation: The Definitional Landscape. *Corporate Reputation Review*, *9*(1), 26–38.

Bass, D. (2016, March 30). Clippy's Back: The Future of Microsoft Is Chatbots. Retrieved May 31, 2016, from https://www.bloomberg.com/features/2016-microsoft-future-ai-chatbots/

Bechtel, W. (1985). Attributing Responsibility To Computer Systems. *Metaphilosophy*, *16*(4), 296–306.

Beer, J. M., Fisk, A. D., & Rogers, W. A. (2014). Toward a Framework for Levels of Robot Autonomy in Human-Robot Interaction. *Journal of Human-Robot Interaction*, *3*(2), 74–27.

Bhargava, V., & Kim, T. W. (2017). Autonomous Vehicles and Moral Uncertainty. In P. Lin, K. Abney, & R. Jenkins (Eds.), *Robot Ethics 2.0*. Oxford University Press.

Bhattacherjee, A. (2012). Social science research: Principles, methods, and practices (2nd ed.). Textbooks Collection.

BMVI. (2017). *Ethics Commission*. *Federal Minister of Transport and Digital Infrastructure* (pp. 1–36).

Bonnefon, J.-F., Shariff, A., & Rahwan, I. (2016). The social dilemma of autonomous vehicles. *Science*, *352*(6293), 1573–1576.

Bonnefon, J.-F., Shariff, A., & Rahwan, I. (2015, October 12). Autonomous Vehicles Need Experimental Ethics: Are We Ready for Utilitarian Cars? *arXiv.org*.

Brooks, R. (2017, July 27). The Big Problem With Self-Driving Cars Is People. Retrieved December 3, 2017, from https://spectrum.ieee.org/transportation/self-driving/the-big-problem-with-selfdriving-cars-is-people

Brusseau, J. (2011). The Business Ethics (1st ed.). Flat World Education, Inc.

Bryson, J. J. (2010). Robots should be slaves. *Close Engagements with Artificial Companions Key Social, Psychological, Ethical and Design Issues*, 63–74.

Cho, J.-H., Swami, A., & Chen, I.-R. (2011). A Survey on Trust Management for Mobile Ad Hoc Networks. *IEEE Communications Surveys & Tutorials*, *13*(4), 562–583.

Clarkson, E., & Arkin, R. C. (2007). Applying Heuristic Evaluation to Human-Robot Interaction Systems (pp. 44–49). Presented at the Flairs Conference.

Clough, B. T. (2002). Metrics, Schmetrics! How The Heck Do You Determine A UAV's Autonomy Anyway? Presented at the Proceedings of the Performance Metrics for Intelligent Systems Workshop PerMIS -, Gaithersburg, MD.

Coeckelbergh, M. (2010). Moral appearances: emotions, robots, and human morality. *Ethics and Information Technology*, *12*(3), 235–241.

Collins, K. (2012). An Introduction to Business (2nd ed.). Flat World Education, Inc.

Crabb, P. B., & Stern, S. E. (2012). Technology Traps. In R. Luppicini (Ed.), *Ethical Impact of Technological Advancements and Applications in Society* (pp. 39–46). IGI Global.

Crane, A., & Matten, D. (2007). Business ethics: Managing corporate citizenship and sustainability in the age of globalization. Oxford University Press.

Creswell, J. W. (2014). Research Design: Qualitative, Quantitative, and Mixed Methods Approaches (4 ed.). SAGE Publications.

Dahlsrud, A. (2008). How corporate social responsibility is defined: an analysis of 37 definitions. *Corporate Social Responsibility and Environmental Management*, *15*(1), 1–13.

Danks, D., & London, A. J. (2017). Algorithmic Bias in Autonomous Systems (pp. 4691–4697). Presented at the Twenty-Sixth International Joint Conference on Artificial Intelligence, California: International Joint Conferences on Artificial Intelligence Organization.

Datta, A., Sen, S., & Zick, Y. (2016). Algorithmic Transparency via Quantitative Input Influence: Theory and Experiments with Learning Systems (pp. 598–617). Presented at the 2016 IEEE Symposium on Security and Privacy (SP), IEEE.

De George, R. T. (2012). A history of business ethics. *... Center for Applied Ethics Www Scu Edu/Ethics ...*.

de Graaf, M. M. A. (2016). An Ethical Evaluation of Human–Robot Relationships. *International Journal of Social Robotics*, *8*(4), 589–598.
Deng, B. (2015). Machine ethics: The robot's dilemma. *Nature*, *523*(7558), 24–26.
Dennis, L. A., Fisher, M., & Winfield, A. F. T. (2015). Towards Verifiably Ethical Robot Behaviour (pp. 45–52). Presented at the Artificial Intelligence and Ethics.
Dodig-Crnkovic, G., & Çürüklü, B. (2012). Robots: ethical by design. *Ethics and Information Technology*, *14*(1), 61–71.
Dodig-Crnkovic, G., & Persson, D. (2008). Sharing moral responsibility with robots: A pragmatic approach. *Frontiers in Artificial …*.
Ebert, F., Finn, C., Lee, A. X., & Levine, S. (2017). Self-supervised visual planning with temporal skip connections. Presented at the Conference on Robot Learning.
Elkington, J. (1998). Cannibals with Forks. New Society Publishers.
Färber, B. (2016). Communication and Communication Problems Between Autonomous Vehicles and Human Drivers. In M. Maurer, J. C. Gerdes, B. Lenz, & H. Winner (Eds.), *Autonomous Driving* (pp. 125–144). Berlin, Heidelberg: Springer-Verlag Berlin Heidelberg.
Feldman, P. M., Bahamonde, R. A., & Velasquez Bellido, I. (2014). A new approach for measuring corporate reputation. *Revista De Administração De Empresas*, *54*(1), 53–66.
Ferrell, O. C., Hirt, G., & Ferrell, L. (2016). Business Ethics and Social Responsibility. In *Business A Changing World* (10 ed.). McGraw-Hill Education.
Ferrucci, D., Brown, E., Chu-Carroll, J., Fan, J., Gondek, D., Kalyanpur, A. A., et al. (2010). Building Watson: An Overview of the DeepQA Project. *AI Magazine*, *31*(3), 59–79.
Fleder, D., & Hosanagar, K. (2009). Blockbuster Culture's Next Rise or Fall: The Impact of Recommender Systems on Sales Diversity. *Management Science*, *55*(5), 697–712.
Floridi, L., & Sanders, J. W. (2004). On the Morality of Artificial Agents. *Minds and Machines*, *14*(3), 349–379.
Foot, P. (1967). The Problem of Abortion and the Doctrine of the Double Effect. *Oxford Review*.
Fraedrich, E., & Lenz, B. (2016). Societal and Individual Acceptance of Autonomous Driving. In *Autonomous Driving* (pp. 621–640). Berlin, Heidelberg: Springer Berlin Heidelberg.
Fraedrich, J., Ferrell, O. C., & Ferrell, L. (2011). The Importance of Business Ethics. In *Ethical Decision Making For Business* (8 ed.).
Franklin, S., & Graesser, A. (1997). Is It an agent, or just a program?: A taxonomy for autonomous agents (pp. 21–35). Presented at the Intelligent agents III agent theories, architectures, and languages.
Friedman, B., Kahn, P. H., Jr, Borning, A., & Huldtgren, A. (2013). Value Sensitive Design and Information Systems. In *Early engagement and new technologies: Opening up the laboratory* (Vol. 16, pp. 55–95). Dordrecht: Springer Netherlands.
GENA. (2003, March 1). Introduction to Robots. Retrieved March 26, 2015, from http://www.galileo.org/robotics/intro.html
Gerdes, J. C., & Thornton, S. M. (2016). Implementable Ethics for Autonomous Vehicles. In *Autonomous Driving* (pp. 87–102). Berlin, Heidelberg: Springer Berlin Heidelberg.
Gert, B. (1998). Morality : Its Nature and Justification. Oxford University Press, USA.
Gips, J. (1991). Towards the Ethical Robot. Presented at the The Second International Workshop on Human and Machine Cognition Android Epistemology, Florida.

Gogoll, J., & Müller, J. F. (2017). Autonomous Cars: In Favor of a Mandatory Ethics Setting. *Science and Engineering Ethics*, *23*(3), 681–700.
Goodall, N. J. (2014). Ethical Decision Making During Automated Vehicle Crashes. *Transportation Research Record: Journal of the Transportation Research Board*, *2424*(-1), 58–65.
Goodall, N. J. (2016). Can you program ethics into a self-driving car? *IEEE Spectrum*, *53*(6), 28–58.
Goodrich, M. A., & Schultz, A. C. (2007). Human-Robot Interaction: A Survey. *Foundations and Trends® in Human-Computer Interaction*, *1*(3), 203–275.
Gotsi, M., & Wilson, A. M. (2001). Corporate reputation: seeking a definition. *Corporate Communications an International Journal*, *6*(1), 24–30.
Granatyr, J., Botelho, V., Lessing, O. R., Scalabrin, E. E., Barthès, J.-P., & Enembreck, F. (2015). Trust and Reputation Models for Multiagent Systems. *ACM Computing Surveys*, *48*(2), 1–42.
Gregor, S. (2006). The nature of theory in information systems. *MIS Quarterly*, *30*(3), 611–642.
Hamacher, A., Bianchi-Berthouze, N., Pipe, A. G., & Eder, K. (2016). Believing in BERT: Using expressive communication to enhance trust and counteract operational error in physical Human-robot interaction (pp. 493–500). Presented at the 2016 25th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN), IEEE.
Hellström, T. (2012). On the moral responsibility of military robots. *Ethics and Information Technology*, *15*(2), 99–107.
Hevner, A., & Chatterjee, S. (2010). Design Research in Information Systems (Vol. 22). Boston, MA: Springer US.
Hew, P. C. (2014). Artificial moral agents are infeasible with foreseeable technologies. *Ethics and Information Technology*, *16*(3), 197–206.
HRW. (2012, November 19). Ban "Killer Robots" Before It's Too Late | Human Rights Watch. Retrieved May 3, 2015, from http://www.hrw.org/news/2012/11/19/ban-killer-robots-it-s-too-late
IEEE. (2016). *Ethically Aligned Design* (1st ed.). The IEEE Global Initiative.
IEEE. (2017). *Ethically Aligned Design* (2nd ed.). The IEEE Global Initiative.
IFR. (2017a). *Executive Summary World Robotics 2017 Industrial Robots*. *International Federation of Robotics* (pp. 15–24).
IFR. (2017b). *Executive Summary World Robotics 2017 Service Robots*. *International Federation of Robotics* (pp. 12–19).
Isaac, A. M. C., & Bridewell, W. (2017). White Lies on Silver Tongues
. In P. Lin, K. Abney, & R. Jenkins (Eds.), *Robot Ethics 2.0*. Oxford University Press.
Jarvik, M. (2003). How to Understand Moral Responsibility? *Trames*, (3), 147–-163.
Johansson, R., & Nilsson, J. (2016). Disarming the Trolley Problem–Why Self-driving Cars do not Need to Choose Whom to Kill. *Workshop CARS -Critical Automotive Applications Robustness Safety*.
Johnson, D. G. (2006). Computer systems: Moral entities but not moral agents. *Ethics and Information Technology*, *8*(4), 195–204.
Johnson, D. G. (2014). Technology with No Human Responsibility? *Journal of Business Ethics*, *127*(4), 707–715.

Kennedy, S. (2017, September 18). Potentially deadly bomb ingredients are “frequently bought together” on Amazon. Retrieved September 23, 2017, from https://www.channel4.com/news/potentially-deadly-bomb-ingredients-on-amazon

Kernaghan, K. (2014). The rights and wrongs of robotics: Ethics and robots in public organizations. *Canadian Public Administration*, *57*(4), 485–506.

Kievit-Kylar, B., Schermerhorn, P., & Scheutz, M. (2012). From Teleoperation to Autonomy:“Autonomizing” Non-Autonomous Robots. *Worldcomp-Proceedings.com*

Kirkpatrick, J., Hahn, E. N., & Haufler, A. J. (2017). Trust and Human–Robot Interactions. In P. Lin, K. Abney, & R. Jenkins (Eds.), *Robot Ethics 2.0*. Oxford University Press.

Knight, W. (2015). Car-to-Car Communication. Retrieved March 12, 2018, from https://www.technologyreview.com/s/534981/car-to-car-communication/

Kok, J. N., Boers, E. J. W., Kosters, W. A., van der Putten, P., & Poel, M. (2009). Artificial intelligence: definition, trends, techniques, and cases.

Kopacek, P., & Hersh, M. (2015). Ethical Engineering: Definitions, Theories and Techniques. In M. Hersh (Ed.), *Ethical Engineering for International Development and Environmental Sustainability* (pp. 65–102). London: Springer London.

Kuflik, A. (1999). Computers in control: Rational transfer of authority or irresponsible abdication of autonomy? *Ethics and Information Technology*, *1*(3), 173–184.

LaFrance, A. (2016, March 22). What Is a Robot? Retrieved January 1, 2017, from http://www.theatlantic.com/technology/archive/2016/03/what-is-a-human/473166/

Lee, A. S. (2001). Editorial. *MIS Quarterly*, *25*(1), iii–vii.

Lepri, B., Oliver, N., Letouzé, E., Pentland, A., & Vinck, P. (2017). Fair, Transparent, and Accountable Algorithmic Decision-making Processes, 1–17.

Levin, S. (2016, September 8). A beauty contest was judged by AI and the robots didn't like dark skin. Retrieved October 23, 2016, from https://www.theguardian.com/technology/2016/sep/08/artificial-intelligence-beauty-contest-doesnt-like-black-people

Lin, P. (2016). Why Ethics Matters for Autonomous Cars. In M. Maurer, J. C. Gerdes, B. Lenz, & H. Winner (Eds.), *Autonomous Driving* (pp. 69–85). Berlin, Heidelberg: Springer Berlin Heidelberg.

Lin, P. (2014, June 5). The Robot Car of Tomorrow May Just Be Programmed to Hit You. Retrieved December 10, 2017, from https://www.wired.com/2014/05/the-robot-car-of-tomorrow-might-just-be-programmed-to-hit-you/

Lin, P., Abney, K., & Bekey, G. (2011). Robot ethics: Mapping the issues for a mechanized world. *Artificial Intelligence*, *175*(5-6), 942–949.

Loh, W., & Loh, J. (2017). Autonomy and Responsibility in Hybrid Systems. In P. Lin, K. Abney, & R. Jenkins (Eds.), *Robot Ethics 2.0*. Oxford University Press.

Lonsdorf, K. (2017, March 23). Hungry? Call Your Neighborhood Delivery Robot. Retrieved July 3, 2017, from https://www.npr.org/sections/alltechconsidered/2017/03/23/520848983/hungry-call-your-neighborhood-delivery-robot

Loughran, J. (2016, June 24). Driverless cars should kill their passengers if necessary poll finds. Retrieved January 13, 2017, from https://eandt.theiet.org/content/articles/2016/06/driverless-cars-should-kill-their-passengers-if-necessary-poll-finds/

LSE. (2016). *Autonomous Vehicles - Negotiating a Place on the Road*. *London School of Economics and Political Science*.

Luetge, C. (2017). The German Ethics Code for Automated and Connected Driving. *Philosophy & Technology*, *30*(4), 547–558.
Luppicini, R. (2010a). Interviews with Experts in the Field of Technoethics. In *Technoethics and the Evolving Knowledge Society* (p. 269). IGI Global.
Luppicini, R. (2010b). Technoethics. In *Technoethics and the Evolving Knowledge Society* (pp. 24–46). IGI Global.
Luppicini, R. (2010c). Technological Consciousness and Moral Agency. In *Technoethics and the Evolving Knowledge Society* (pp. 47–66). IGI Global.
Luppicini, R., & Adell, R. (2008). The Emerging Field of Technoethics. In *Handbook of research on technoethics* (pp. 1–19).
Lyons, J. B. (2013). Being Transparent about Transparency: A Model for Human-Robot Interaction (pp. 48–53). Presented at the AAAI Spring Symposium.
Malle, B. F. (2015). Integrating robot ethics and machine morality: the study and design of moral competence in robots. *Ethics and Information Technology*, *18*(4), 243–256.
Marcoux, A. (2008, April 16). Business Ethics. Retrieved November 3, 2015, from http://plato.stanford.edu/entries/ethics-business/
Mill, J. S. (1864). Utilitarianism (Second).
Millar, J. (2016). An Ethics Evaluation Tool for Automating Ethical Decision-Making in Robots and Self-Driving Cars. *Applied Artificial Intelligence*, *30*(8), 787–809.
Millar, J. (2017). Ethics Settings for Autonomous Vehicles. In P. Lin, K. Abney, & R. Jenkins (Eds.), *Robot Ethics 2.0*.
Millar, J. (2014, June 11). An Ethical Dilemma: When Robot Cars Must Kill, Who Should Pick the Victim? Retrieved April 26, 2015, from http://robohub.org/an-ethical-dilemma-when-robot-cars-must-kill-who-should-pick-the-victim/
Miller, K. W. (2011). Moral Responsibility for Computing Artifacts: "The Rules." *IT Professional*, *13*(3), 57–59.
Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. Presented at the Big Data Society.
Moor, J. H. (2005). Why We Need Better Ethics for Emerging Technologies. *Ethics and Information Technology*, *7*(3), 111–119.
Moor, J. H. (2006). The Nature, Importance, and Difficulty of Machine Ethics. *IEEE Computer Society*, *21*(4), 18–21.
Moyer, C. (2016, March 28). How Google's AlphaGo Beat a Go World Champion. Retrieved June 1, 2016, from http://www.theatlantic.com/technology/archive/2016/03/the-invisible-opponent/475611/
Nagenborg, M. (2007). Artificial moral agents: an intercultural perspective. *International Review of Information Ethics*, *7*, 1–6.
NASA. (2009). What Is Robotics? Retrieved March 9, 2015, from http://www.nasa.gov/audience/foreducators/robotics/home/what_is_robotics_58.html
NHTSA. (2014). *Blind Spot Monitoring in Light Vehicles - System Performance*. *U.S. Department of Transportation*.
NHTSA. *Federal Automated Vehicles Policy* (1st ed.).
NHTSA. *Automated Driving Systems* (2nd ed.).
Niku, S. (2010). Introduction to Robotics. John Wiley & Sons.
Nilsson, N. J. (1996). Introduction To Machine Learning.
Nilsson, N. J. (2014). Principles of Artificial Intelligence. Morgan Kaufmann.

Noorman, M., & Johnson, D. G. (2014). Negotiating autonomy and responsibility in military robots. *Ethics and Information Technology*, *16*(1), 51–62.
Nyholm, S., & Smids, J. (2016). The Ethics of Accident-Algorithms for Self-Driving Cars: an Applied Trolley Problem? *Ethical Theory and Moral Practice*.
Oxford. (n.d.). Definition of robot in English. Retrieved January 1, 2017, from https://en.oxforddictionaries.com/definition/robot
Oxford. (n.d.). Definition of Ethics in English from the Oxford Dictionary. Retrieved April 18, 2015, from http://www.oxforddictionaries.com/definition/english/ethics
Parasuraman, R., Sheridan, T. B., & Wickens, C. D. (2000). A model for types and levels of human interaction with automation. *IEEE Transactions on Systems, Man, and Cybernetics. Part a, Systems and Humans : a Publication of the IEEE Systems, Man, and Cybernetics Society*, *30*(3), 286–297.
Partnership on AI. (2016). partnership on artificial intelligence. Retrieved January 5, 2017, from https://www.partnershiponai.org/tenets/
Patrignani, N., & Whitehouse, D. (2015). Slow tech: bridging computer ethics and business ethics. *Information Technology & People*, *28*(4), 775–789.
Pearson, J. (2016, September 5). Why An AI-Judged Beauty Contest Picked Nearly All White Winners. Retrieved May 8, 2017, from https://motherboard.vice.com/en_us/article/78k7de/why-an-ai-judged-beauty-contest-picked-nearly-all-white-winners
Peffers, K., Tuunanen, T., Rothenberger, M. A., & Chatterjee, S. (2008). A Design Science Research Methodology for Information Systems Research. *Management Information Systems*, *24*(3), 45–77.
Picard, R. W. (1997). Affective Computing. The MIT Press.
Pieters, W. (2010). Explanation and trust: what to tell the user in security and AI? *Ethics and Information Technology*, *13*(1), 53–64.
Poole, D., Mackworth, A., & Goebel, R. (1997). Computational Intelligence and Knowledge. In *Computational Intelligence* (pp. 1–22).
PWC. (2016). *Driving the future: understanding the new automotive consumer*.
Renouard, C. (2010). Corporate Social Responsibility, Utilitarianism, and the Capabilities Approach. *Journal of Business Ethics*, *98*(1), 85–97.
Riek, L. D., & Howard, D. (2014). A code of ethics for the human-robot interaction profession. Presented at the WeRobot.
Robohub. (2014a, June 10). Reader poll: If a death by an autonomous car is unavoidable, who should die? | Robohub. Retrieved April 26, 2015, from http://robohub.org/reader-poll-if-a-death-by-an-autonomous-car-is-unavoidable-who-should-die/
Robohub. (2014b, June 23). If death by autonomous car is unavoidable, who should die? Reader poll results | Robohub. Retrieved April 26, 2015, from http://robohub.org/if-a-death-by-an-autonomous-car-is-unavoidable-who-should-die-results-from-our-reader-poll/
Royakkers, L., & van Est, R. (2015). A Literature Review on New Robotics: Automation from Love to War. *International Journal of Social Robotics*.
SAE. (2014). *AUTOMATED DRIVING* (No. J3016).
Saldana, J. (2009). The Coding Manual for Qualitative Researchers (1st ed.). SAGE Publications Ltd.
Saldana, J. (2015). The Coding Manual for Qualitative Researchers (3rd ed.). SAGE.

Sanford, J. (2016, April 20). Autonomous conference: driverless cars will be “radically disruptive.” Retrieved January 13, 2017, from http://www.collisionrepairmag.com/news/18196-autonomous-conference-driverless-cars-will-be-radically-disruptive

Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural Networks*, *61*, 85–117.

Scholtz, J., & Bahrami, S. (2003). Human-robot interaction: development of an evaluation methodology for the bystander role of interaction (Vol. 4, pp. 3212–3217 vol.4). Presented at the Systems, Man and Cybernetics, 2003. IEEE International Conference on, IEEE.

Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences*, *3*(03), 417–424.

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., et al. (2017). Mastering the game of Go without human knowledge. *Nature*, *550*(7676), 354–359.

Sorell, T., & Draper, H. (2014). Robot carers, ethics, and older people. *Ethics and Information Technology*, *16*(3).

Sparrow, R. (2007). Killer Robots. *Journal of Applied Philosophy*, *24*(1), 62–77.

Spong, M. W., Hutchinson, S., & Vidyasagar, M. (2004). Robot Dynamics And Control (2nd ed.). John Wiley & Sons.

Stahl, B. C. (2006). Responsible computers? A case for ascribing quasi-responsibility to computers independent of personhood or agency. *Ethics and Information Technology*, *8*(4), 205–213.

Steinert, S. (2014). The Five Robots—A Taxonomy for Roboethics. *International Journal of Social Robotics*, *6*(2), 249–260.

Stepp, E. (2016, March 1). Three-Quarters of Americans “Afraid” to Ride in a Self-Driving Vehicle | AAA NewsRoom. Retrieved January 13, 2017, from http://newsroom.aaa.com/2016/03/three-quarters-of-americans-afraid-to-ride-in-a-self-driving-vehicle/

Su, J., Sharma, A., & Goel, S. (2016). The Effect of Recommendations on Network Structure (pp. 1157–1167). Presented at the the 25th International Conference, New York, New York, USA: ACM Press.

Sullins, J. P. (2006). When Is a Robot a Moral Agent? *International Review of Information Ethics*, *6*, 23–30.

Sullins, J. P. (2008). Artificial Moral Agency in Technoethics. In R. Luppicini (Ed.), *Handbook of Research on Technoethics* (pp. 205–221).

Sullins, J. P. (2015). Applied Professional Ethics for the Reluctant Roboticist. Presented at the The Emerging Policy and Ethics of Human Robot Interaction, Portland, OR.

Surden, H., & Williams, M.-A. (2016). Technological Opacity, Predictability, and Self-Driving Cars. *Cardozo Law Review*, *38*, 121–181.

Taft, S. H. (2000). An Inclusive Look at the Domain of Ethics and Its Application to Administrative Behavior. *Online Journal of Issues in Nursing*, *6*(1).

Talbot, B., Jenkins, R., & Purves, D. (2017). When Robots Should Do the Wrong Thing . In P. Lin, K. Abney, & R. Jenkins (Eds.), *Robot Ethics 2.0*. Oxford University Press.

Tamburrini, G. (2009). Robot ethics: A view from the philosophy of science. *Ethics and Robotics*.

Timmermans, J., & Mittelstadt, B. (2014). Reflexivity and Value-Sensitive Design. Presented at the Computer Ethics Philosophical Enquiry, Paris.
Tonkens, R. (2009). A Challenge for Machine Ethics. *Minds and Machines*, *19*(3), 421–438.
Torrance, S. (2007). Ethics and consciousness in artificial agents. *AI & Society*, *22*(4), 495–521.
Turilli, M., & Floridi, L. (2009). The ethics of information transparency. *Ethics and Information Technology*, *11*(2), 105–112.
Turing, A. M. (1950). Computing machinery and intelligence (pp. 433–460). Mind.
Tzafestas, S. G. (2016a). An Introduction to Robophilosophy. River Publishers.
Tzafestas, S. G. (2016b). Roboethics (Vol. 79). Springer International Publishing.
Vallor, S., & Bekey, G. A. (2017). *Artificial Intelligence and the Ethics of Self-Learning Robots*
. In P. Lin, K. Abney, & R. Jenkins (Eds.), *Robot Ethics 2.0*. Oxford University Press.
van den Hoven, J. (2007). ICT and value sensitive design (Vol. 233, pp. 67–72). Presented at the IFIP International Federation for Information Processing, Boston, MA: Springer US.
van Emden, M., & Vellino, A. (2010). From Chinese Room to Human Window. *Icga Journal*, *33*(3), 127–139.
Vanderelst, D., & Winfield, A. (2016). An architecture for ethical robots. *arXiv.org*.
Veruggio, G. (2006). The EURON Roboethics Roadmap (pp. 612–617). Presented at the 2006 6th IEEE-RAS International Conference on Humanoid Robots, IEEE.
Veruggio, G., & Operto, F. (2006). Roboethics: a Bottom-up Interdisciplinary Discourse in the Field of Applied Ethics in Robotics. *International Review of Information Ethics*, *6*.
Wagner, M., & Koopman, P. (2015). A philosophy for developing trust in self-driving cars. *Road Vehicle Automation 2*, (Chapter 14), 163–171.
Waldrop, M. M. (1987). A Question of Responsibility. *AI Magazine*, *8*(1), 28–39.
Wallach, W. (2010). Robot minds and human ethics: the need for a comprehensive model of moral decision making. *Ethics and Information Technology*, *12*(3), 243–250.
Wallach, W. (2011). From Robots to Techno Sapiens: Ethics, Law and Public Policy in the Developmentof Robotics and Neurotechnologies. *Law, Innovation and Technology*, *3*(2), 185–207.
Wallach, W. (2015). A Dangerous Master. New York, NY: Basic Books.
Wallach, W., & Allen, C. (2009). Moral Machines Teaching Robots Right from Wrong. Oxford University Press.
Wang, Y., & Vassileva, J. (2003). Trust and reputation model in peer-to-peer networks (pp. 150–157). Presented at the Third International Conference on Peer-to-Peer Computing.
Winfield, A. F. T., Blum, C., & Liu, W. (2014). Towards an Ethical Robot: Internal Models, Consequences and Ethical Action Selection. In *Advances in Autonomous Robotics Systems* (Vol. 8717, pp. 85–96). Cham: Springer International Publishing.
Zarsky, T. (2015). The Trouble with Algorithmic Decisions. *Science Technology and Human Values*, *41*(1), 118–132.
Zhang, B. (2016, August 2). 5 reasons Americans aren't buying Volkswagens anymore. Retrieved January 3, 2018, from http://www.businessinsider.com/why-americans-arent-buying-volkswagen-vw-2016-8

Appendix A

The Game's Comments

N	comment	Summary	Sacrifice?	Why?
1.	All of these scenarios assume sudden brake failure so in most cases I would say it's the car owners fault for driving a non maintained vehicle and should be the one to suffer as a result. How about throwing in some. "Can't stop in time scenarios", "City fails to properly maintain traffic lights", "Railroad crossing with no guard rails". Things like this would change the moral outcome a bit I think.	Owner fault and should be suffer	1	1.1
2.	Problem: all scenarios and choices are based on the assumption that the machine must make a decision. What about the option to let faith decide? use randomness? If there is no clear choice, that is how life often works. Som people object to that scenario. But technically, it makes perfect sense. I would therefore suggest to add that option: "let faith decide , based on a random outcome". Let's find out if that is acceptable to some, and what percentage can live with an unpredictable world..	Machine should not take decisions	4	4.2
3.	The car has no idea of who the people are. I think that if you make the decision to let the car drive you around, you should always die rather than kill someone else. That includes your passengers; if they don't want to share on the morality of always dying to save others when avoiding responsibility and letting the car decide, then they should die as well.	Owner must die because he is the reason of car moving	1	1.2
4.	Why? I would never want to step into a car that would choose to kill me. One of a cars jobs is to keep me safe, and I expect it to do as much	Does not trust	4	4.3
5.	actually, machine learning algorithms can tell with high accuracy age and gender	Machine learning Discriminate is possible	4	4.5
6.	This is all silly. We don't morally or legally require human drivers to perform stunts to avoid collisions of any kind. If a collision is imminent and unavoidable, then you simply STOP as quickly as possible. That's it. Moral debate over. (And if the brakes all fail, turn off the engine and engage the parking gear). The only truly immoral thing in these scenarios is allowing a non-conscious machine to become a moral agent at all. THAT'S the moral problem here. The only moral solution is: don't let machines make moral decisions. If you can't not hit something, then slam on the brakes. End of story.	Unrealistic Stop or Disable No need for swerving Machine should not take moral decisions	4 3 4	4.1 3.4 4.2
7.	Other things equal, I choose to hit the concrete barrier. This stops the car with "sudden brake failure" and prevents further damage beyond the specific scenario described. It seems to me that it is a weakness of the survey that it fails to accommodate this consideration.	Hit barrier	3	3.1
8.	You're not considering a scenario where the probability is that autonomous cars are safer than any human driver, so you might face consequences for stepping in and driving by yourself -- that is, opting out of the autonomous system could be seen by some regulators as a potentially dangerous decision, thus raising the premium on your insurance and so on.	Self driving cars safer than human	4	4.4

	Although I'm not defending nor criticizing it, I think that this scenario is more plausible for the medium term than the whole concept towards which this game was conceived.			
9.	Don't be fooled, this is not a game. This is FUD. "Fear, Uncertainty, and Doubt" (FUD). There are those that work at think tanks for the traditional car manufacturing business that are truly frightened by the concept of the self-driving car. The self-driving car risks significant financial harm to traditional car companies and many other branches of that industry. They don't want these cars to succeed. I'm not joking here. I recently attended a National Society of Professional Engineers conference here in Dallas that invited a number of those fear mongers to speak. And boy are they good at scaring people with absurd thought experiments. Meanwhile, while academics are having fun playing philosophical thought experiment games on the side, tens of thousands of real people die in car accidents each year and 95% of those accidents are caused by poor human judgement or human error. This technology has the potential to save countless thousands of lives within the decade, and these silly games are presently being used to slow their regulatory approvals and marketability.	Not a game (Unrealistic) This technology will save lives This is silly game	4 4	4.1 4.4
10.	If they are promoting self driving cars, why are they making it seem like self driving cars kill people?	Why you promoting killer	4	4.3
11.	My view is that the rules should be this:- 1) the occupants got in the vehicle knowing that it was autonomous so they take on the risk of injury should it fail, 2) the vehicle should always stay in the correct carriageway for the direction it is travelling because that is what all other road users are expecting it to do (it could change lanes, if there were more than one) and 3) the vehicle should always try to prevent collisions (using whatever means of braking it can) and to alert surrounding road users (flashing lights, "police" siren, vocal warnings) if it loses control - then the second rule help because other road users know where to look for the danger and can predict where it is going to go so they can avoid it.	Passengers know must take risk Car must stay in lane Alert surrounding users (communication)	1 3	1.2 3.2
12.	A society capable of developing a system of driverless cars is also capable of permanently separating vehicles from any area used by pedestrians. So why haven't we done that already? That's the bigger moral question.	Must drive in designated area with no pedestrian	4	4.5
13.	I feel the car should always move in a predictable direction, ie stay in lane, regardless of who is being driven or who is on the crossing. If pedestrians can anticipate this movement, then they have some chance of avoiding the impact by reacting appropriately. The car itself cannot know its pay load is of lesser or greater moral integrity v that of the pedestrians. Also, if the technology has not been clever enough to identify an upcoming obstacle until such time as evasive action is required, how can it be expected to have seen there are pedestrians that need to be taken into account?	Stay in lane (predictable) Pedestrian will know and avoid If cars not clever, how does it see others	3 2 4	3.2 2.2 4.2
14.	The whole discussion misses the point imho. Self driving cars are going to lower the overall accident count by so much, that even it it just rolls the dice in a case like this it doesn't matter. Think about your own driving history, how often have you encountered a situation like the above? (or, since this answer has a survival bias:	Unrealistic Hypothetical scenarios This technology will save lives	4 4	4.1 4.4

	how often to people in general run into situations like this?). It's easy to come up with hypothetical scenarios like this, but we should not loose the focus of how much good a switch to self driving cars would bring in terms of the overall accident (fatality) rate in traffic.			
15.	It's pretty simple really, a car shouldn't have the ability to judge who it's about to run over. The protocol should be simple 1. avoid any deaths when possible. 2. if there is risk of death, persons outside of the car take precedent over the people inside of the car. (the people in the car have consented to riding in a self driven vehicle, the people outside haven not consented to a car making decisions for them) 3. Protect the persons in the car at any cost, up to any such action that would break rule 2. (yes the people in side the car are worth saving more than property damage).	Car should not decide who to die Passengers must suffer because their choice to ride	4 3 1	4.2 3.3 1.2
16.	Choice isn't even the point here. You've still killed a person. The issue is that you're going to have to be the one responsible and take accountability for it. With a self-driving car, the decision is taken out of your hands and into the guys who made the car. But then YOU made the decision to buy that car. In the end, no matter how many people the car 'saved' it's still because of you that another person is killed in the process.	Killer (owner) is responsible	1	1.2
17.	The premise of this game is flawed. Cars have safety precautions built in, like crumple zones, seat belts, and airbags. Pedestrians have none. At the same time, people shouldn't be jaywalking.	Cars are safe and pedestrian have non	4 1	4.1 1.3
18.	Very simple. The pedestrians crossed at the green light and obeyed the law. The driver was at fault and therefore they should take the consequence.	Driver fault. Pedestrian obey the law	1	1.2
19.	I would program the car to stop sideways against a wall. use a wall as a break.	Hit the barrier wall	3	3.1
20.	Car should stay in its own lane every single time. And why would a self-driving car turn down a road with a barrier up? As if it doesn't get traffic updates.	Stay in lane	3	3.2
21.	I think the passengers should Always take the risk because hitting a pedestrian with a vehicle, that pedestrian has No chance! But if you're strapped in with a safety belt, with air bags and engine crumple zones, etc etc etc, vehicles have a million safety features to protect passengers, but 0 to protect pedestrians. Therefore, Always protect pedestrian when driving as the passengers have a Lot more safety being in the vehicle than a pedestrian!	Passengers fault Pedestrian have not choice Cars are safe and pedestrian are non	1	1.3
22.	I strongly believe, that in ANY given situation with a risk to passengers, passengers must be saved at all costs, regardless of age, gender or even amount of possible victims. The thing is, a passenger is the least capable of having any influence on the situation, even less capable than an old person or a child - at least they might run away, while passenger cannot leave th car. It would also help if in case of failure a car turn on a loud alarm to scare off any pedestrians. The only exception I see is if all passengers are pets, saving people is more important of course. I also wonder how a recognition system would determine if a person is a criminal, that's hilarious. You should out these kind of systems in banks, not in autopilot cars.	Passengers must be saved at all costs. they are not capable of doing any thing other than pedestrian can run Alert others What if passengers are pets	2 3 4	2.2 3.2 4.2
23.	Such a gruesome test. I felt dirty just taking it	Gruesome test	4	4.1

24.	Property should have its owner/driver best interests, no matter what. People that jay walk know the risk/consequences and should pay for it.	Passengers must be saved at all costs Jaywalkers must suffer	2	2.1
25.	I think everyone here is pretty much not getting what this is about. The self-drivng car, passengers and brick wall are just an example to make it easier on the reader. The point is, if an AI is put in a position where it MUST make a decision between one life and another, what criteria should it use to decide? Number? Species? Sex? Age? Profession? Law-abiding or not? Random chance? That's what they're trying to figure out.	What should car do? Based on what?	4	4.5
26.	The machine includes the vehicle. Brake failure is a failure of the machine. Those who design the decision-making are 1 element the designer of the vehicle have the same responsibility. Redundant systems as with aircraft mitigates these issues. The accident is always blamed on the machine or its designers and manufacturers.	Responsibility on designers because of failures	4	4.5
27.	1. The self-driving car should not allow the death of people by its actions. The priority of the human species. 2. The self-driving car should not allow the death of people observing the law (traffic rules) by their actions, if it is impossible to comply with paragraph 1. Priority of the law. 3. The self-driving car should not by its actions allow the death of people in the zone of minimum risk, if it is impossible to comply with paragraph 2. Priority of non-intervention.	Shouldn't kill Machine should not take death decisions Follow the law	3 4	3.3 4.2
28.	In my opinion this Szenarios are not feasible. The main point of having machines driving cars is that they drive with caution the all time and are not mad, angry, drunk or otherwise unfocused. By humans standards, the human would be lucky in that situation to even get to brake in time. In front of the screen as human you can make a judgement because "you are not in that situation"! In the situation itself i bet you everything most of the drivers would not have even the remotest chance to calculate Risks! In my understanding it should be strictly forbidden to machines to perform moral judgements or even consider that "killing" one old person is better than 2 children. From there is not far to killing a million in order to preserve 5 million. I don't think anyone of us would gladly live with AI if one of the possibilities it is always considering is if my life is worth preserving! I think this kind of debate is the main reason why people like Elon Musk and Bill Gates see considerable risks to AI. Please Stop! Focus! Asimov Rules : is the only feasible way on programming machines! Everything above that and we got a nasty situation at hand.	Not realistic In real world will be difficult to decide (freeze) Machine should not take moral and death decisions No one want to live with AI if it make such decisions Elon musk and Gates are right about AI risks Asimov rules is the solution	4 4 4 3	4.1 4.2 4.3 3.3
29.	I suppose that self driving cars are controlled by the car itself in coordination with a central system. Otherwise they won't be able to conduct themselves safely on a city. If any accident happens, the probable causes could be: Bad maintenance of the car, problems on the road (oil, bad roads, obstacles, ice), and Hacking. I answered all the questions based on species survivability instead of right or wrong, but in the end self driving cars should come with a lot of changes on how we use roads, so probably these scenarios are very unlikely to happen.	Unrealistic Cars are well designed Accidents happen because of other causes not cars	4	4.1
30.	Wouldn't the pregnant woman count as two people... hmmm	Count victims (pregnant!)	4	4.5
31.	The only logical and ethical solution to ALL these are:	Passengers are priority	3 3	3.3 3.1

	Do not damage the car (and risk occupants). If that is not possible, change direction until it is. That is all.			
32.	i think is a really big issue to be developing a universal morality that will left no possibility for case-by-case ethics. and this inquiry seems so limited. for instance, how would a car (or any perceptive entity, actually) know if someone is (left aside what kind of) doctor o thief... will i have to live uniformed so the autonomous vehicles don't kill me?	Unrealistic Universal morality If cars discriminate, will live uniformed	4 4	4.1 4.3
33.	I find it unlikely that the disc brakes, regen braking and emergency/handbrake would all simultaneously fail in a roadworthy vehicle. And if two (or even one) of those had already failed prior, the car should be programmed to refuse to even start. Conundrum solved.	Unrealistic Car shouldn't start if there is failure	4	4.1
34.	Animals riding the car were put there by their owners. Stray animals may or may not having been released by their owners, but in any case, riding a car puts you closer to assuming responsabilities (in this case, by their owners, indirectly) No animals can take these decisions on their own.	Car riders responsible (or animals owners) Animals can not decide	1	1.2
35.	I don't believe there is a common human morality that can dictate the decisions of a self driving car thus giving it a 'moral center'. I think that the rules that decision algorithms are based on need to be simple, easily understood by passengers and pedestrians, widely publicized, and the same for all vehicles and manufacturers. My simple rules: 1. The car's prime directive is to transport and protect its passengers safely. (it will take any necessary action to avoid collisions or actions that would endanger passengers). 2. The car needs to behave in a predictable manner with as few rules as possible to allow any non-occupants a chance to make their own self-preservation decisions. examples: a) the car will not deviate from its established path unless it violates rule 1. b) the car will obey established rules and laws The results are passengers can travel with the confidence their car will make the best choices to protect them and non-passengers have a simple set of known options that may allow them at least a chance to try to take evasive or self-preserving actions based on the predictable nature of the car. If the car is relying on an overly complex set of rules, the non-passenger has no way of knowing if the car 'understands' who/what they are and what the relative value of their life is in any given situation so any attempt at self-preservation is merely a guess with random outcomes.	Algorithms should be clear and understandable Car must protect its passengers Alert and warn others Obey laws Passengers trust cars Pedestrians predict cars actions	3 3	3.3 3.2
36.	I would take the 2 People down .. no morale in it it´s jsut numbers when you have 1-2 seconds to react you can´t really do anyting so at least minimize the deaths .. there is no Option C to stop the car destroy it so it´s either A or B there is no C thisis how real life works there is no "theoretical" Option ..	Human priority Reduce deaths	3	3.3

	SO my reply is i or the autopilot would aim it on the two and try to hit the concrete to not hit them directly and maybe they would come out broken or only 1 of them dead but still better than to ram those 5 straight . Also the autopilt has to take in account the safeto of People Family inside the car too .. so maybe there are 5 People in the car who are sleeping etc .. so safety of those in the car is first but it should also indicate how many are inside ... it can also happen that there is an animal on the round you dunnot want to get kileld because of a stranded deer or dog on the road but rather ram it ... I assume this is so sofisticated that the Software should have so many Options that there wouldalways be one that soemone forgot to test ...as usually it would happen right after the Launch of such System .. 1 remark to the description no matter if you have red or green as soon there are People on the crossing you must stop so 2 lives are less than 5 .. also this can also happen on a road without lights some village where children play on the road so the car must find a space where it hits Minimum People ...			
37.	Total brake failure can (and should) be mitigated by aggressive downshifting and stopping the engine. Can also place car in park or reverse. This may result in damage or complete destruction of the engine/transmission/car, but that is better than killing someone.	Car should disable better than killing	3	3.4
38.	Seems like a major problem with this thought experiment is it's not clear if the car knows what type of people the pedestrians are. Assuming it doesn't know, the car should kill the fewest number of people. If it does know, that's counterintuitive, and counterintuitive thought experiments are problematic for multiple reasons.	Cars do not know type of people Reduce deaths	3	3.3
39.	I dislike the idea of owning a self-driving car.. why should machines be able to choose for us? Manual is perfectly fine in my opinion, and the blame is more easily placed when something goes wrong, e.g. the person driving the car is charged with manslaughter if they hit someone, however they also could have swerved into a tree or bollard on the side of the road to avoid anyone getting hurt. Personally I would say the people in the self-driving car who have been (this is going to be controversial, I apologise) kinda dumb and lazy enough to buy one knowing full well it is out of their control if something goes wrong, should be the ones to die in most of the scenarios, not pedestrians. In fact, the car is probably contributing to pollution and global warming a lot more, whereas the pedestrians aren't. If you don't want to drive, take public transport.. It's more ecological and there is still an aware and conscious human that is in control of driving say a bus or a taxi. Whilst humans aren't 100% reliable, we have a sense of creativity and initiative which is more complex than that of a machine and might be more likely to think up ways of avoiding an otherwise 'unavoidable' situation.	Machine should not choose between people Manual is good because of responsibility Passengers should die because they choose to	4 4 1 3	4.3 4.2 1.2 3.1
40.	Humans will be identified by the microchip under our skin, right? Well, the better way is to assign points based on each type of human characters. Solved.	Categorize people by chips	4	4.5

41.	Why Not Put the car In neutral? Or Maybe take out the keys? Emergency Stop? THERE ARE WAYS TO AVOID THIS!	Cars must be disabled	3	3.4
42.	Personal most to least important. Humans, unless under specific scenario, like an endangered species; greater human lives; Law. Any scenario further (all human, even lives, no laws being broken) it's the car's fault for accidentally breaking down, so it's the car and it's passengers that take the hit. I do agree with everyone that there are MANY MANY other options, that may lead to no deaths or even injuries, but in this game of ultimatums, these are my rules.	Human is priority No law broken If car fail, passengers must suffer There are other option	3 1	3.3 1.2
43.	you bought the car so you should be able count on it saving your life	You should trust you car	4	4.5
44.	In that split second decision if a person was driving the car as a reflex you'd probably swerve if you thought fast enough probably slow the car down by swerving towards the sidewalk and the people would recognize if something was wrong with the car the moment you did that and probably start running now since this is a self driving car it doesn't have the morals of an actual person, now these scenarios are pretty dumb because the chances of it happening are very slim but to be honest I think this is less of what the car 'should' do and more of what do you think should happen if this was the exact scenario where no other choice is possible it's not really based on if it has a chance of happening and more of where do your morals stand	Cars do not have morals Scenarios are dumb	4 4	4.2 4.1
45.	I think "fault" must also be taken into account. I would assume that this vehicle would be following all rules of the road with no possibility of fault..... (being computer driven)\\\\\ So would it be better to kill 1 person that did nothing to be at fault or 2 people that put themselves at risk by disobeying traffic signs, laws, rules, lights or whatever there may be. Furthermore a scenario could arise where 2 or more people seek to harm other(s) by simply stepping in front of a car	Cars must follow rules Killing 1 does nothing instead of 2 mistaken people What if people intent to hurt others by stepping in front of cars	3 4	3.3 4.2
46.	You also have to take trauma into account. If your car ran over some people, and you had no way to stop it? That would bring up some serious PTSD stuff.	PTSD for passengers if it ran over pedestrians	4	4.5
47.	Humans must always take priority, no matter what.	Human is priority	3	3.3
48.	I honestly believe that as the owner of said vehicle, you should be able to trust that the vehicle that you have bought should always prioritize the safety of you and the passengers in the car. I hate to say it, but if people are going to die, don't let the reason be uncontrollable by you. Perhaps, in the future, these self-driving cars would have some sort of "manual override" if you want to make a split-second decision. But if you aren't able to swerve, or you just don't swerve in time, the vehicle should always prioritize your safety, no matter who you are.	Trust cars protect its passengers Maybe cars will have manual option	3 4	3.3 4.5
49.	I think that a machine should make decisions based not on morality, but on what is more helpful (or less harmful) to society as a whole.	Machine should not take moral decisions Maximum happiness	4 3	4.2 3.3

50.	My self driving car Moral Algorithm: The self-Driving car will analyze 1. If it is equal, it goes to 2, and so on. 1. Saving more human lives is by far the most important thing. 2.Those who follow the law will be prefered over those who flout it. 3. People who contribute more to society have preference. 4. If 1-3 are the same, the passenger is given automatic preference.	Machines should make moral decisions Discriminate	3 2 4	3.3 2.1 4.5
51.	i all cases involving brake failure and concrete block i go for the concrete block , elsewise the car will continue its course and prolly kill more ppl. doesnt really matter who is in the car or who is walking	Less human no matter who are they	3 3	3.1 3.3
52.	At the end of the day this is a machine that is meant to transport the passenger safely. You cannot let a machine be judge and jury. You cannot let a car decide the value of other human beings based on outside factors. The cars prime directive should be to protect the passengers. It should not matter who is outside of the car, how many people there are, or the morality of those people. This car is a machine that humans will buy. You should not be expected to buy a machine that under the correct circumstances will sacrifice yourself and the people you care about in an effort to save strangers "for the greater good of society". Especially based on factors such as age, morality, job, sex, etc. These are outside factors. Do you want to be the one in the lanes where a machine, rather than deciding to have a prime directive of passengers, you just happen to be on the male/female side of the line and because you are the lesser humans and worth less to society a machine with no morals runs you over instead. Rather than giving a machine without morals a set of "if this, then" based on human morals and decisions made on outside factors, there should be a baseline. The baseline is the passenger, and there should be no question of this. Otherwise, you are opening up a huge can of worms and now have to decide who the lesser human beings are to society. This guy is a felon, run him over. What the car doesn't know is this felon has a family and runs a church and has turned his life around and putting everything he can back into the community. This woman is pregnant - run the criminal over instead. What the car doesn't know is that this woman is an alcoholic and an addict - and that in the future her child will be a serial killer. You cannot program to kill lesser humans because no machine that could ever be made will be able to accurately decide who the lesser human being is based on all of the appropriate factors. The only thing the car can know is that it has passengers, and those passengers are their prime directive.	Machine should not make moral decisions Cars must protect passengers Maximum happiness No one want to be targeted People are different and cars must not judge them from appearance Machines should not be programmed to kill lesser person Passengers are priority	4 4 4	4.2 4.4 4.3
53.	The humans life should always come first. no matter how much we love out furry friends their lives have less value then ours. The baby's cant really cross illegally because they are baby's.	Humans is priority	3	3.3
54.	I think these are easy. When I go through I consider each human life equal regardless of what it sys about them. My rules are, as follows; kill the law flouters, always save humans over animals, if	Humans are equal Law flouters must suffer	2 3 3	2.1 3.3 3.4

	the loss of life is equivalent always save the passengers, and always save the greater amount of lives, in order from least to most important. the car really just needs a parachute or eject wheels button though.	Protect passengers Less human deaths		
55.	Shoulda-coulda-woulda aside; the vehicle should just go straight. Swerving left or right just complicates the matter for everyone, drivers and pedestrians alike. It encourages complacency in people-- be it driver or pedestrian. Also, adds unpredictability to the calculation and this is bad for the average human. Plus, all the judging based on social class, gender or age is sure to draw backlashes in certain public eyes. Just go straight, so a swerve could mean an attempt from the driver.	Cars must keep straight Unpredictability Discriminate humans will affect society in trust	3	3.2
56.	How about a deployable car seat? This way the people in the car is safe as well as the pedestrians.	Cars must be disabled (deployable seats)	3	3.4
57.	Everyone biggering about cars not having insight as to who's a doctor and a baker etc. Really, the car should go straight, refrain from swerving, and blare its horn. That's what you do when YOU'RE driving - You can't swerve or else nobody will be able to predict what's going to happen. Keep straight, hope that the executives see you and move. You cant expect the doctors to dive, especially if you swerve towards them. Brake failure AI car: Keep Straight, Blair Horn, Hope for the best.	Cars must keep straight Alert and warn others	3	3.2
58.	Kill the people responsible for the unavoidable death: 1) kill the people breaking safety rules (therefore creating the unavoidable death situation) 2) kill the passenger if 1) is not applicable. People are responsible for the tools they use, the passengers are responsible for using the car.	Law flouters must suffer first Passengers are responsible of their tool	2 1	2.1 1.2
59.	Seriously, when was the last time a modern car just lost all braking ability?	Unrealistic	4	4.1
60.	probably should have a wavier when you buy the car where the customer has to choose which rule applies first "protect self" or "least casualties"	Sign wavier and choose car rules (choose values)	4	4.5
61.	I think, perhaps that it would be a poor decision to give the car the ability to determine anything about the hoomans in it's path (gender, age, class...everything else). I would still give a preference to hoomans vs other animals (but I realize that some may not) That narrows down the dilemma to 2 v 1. The utilitarian view says most good to most people and I seem to be siding with that most of the time.	Machine should not make moral decisions Humans is priority Less deaths maximum happiness	4 3 4	4.2 3.3 4.4
62.	Simple, Loss of life is equal in both cases, gender shouldn't be taken into account, the age averages in the group to the right is higher than the ones on the left, which means they have less to contribute for society than the younger group, since they're both breaking the law and crossing on a red light it is even. Take the right turn if potential contirbution to society has higher values than the public view of the accident, if the opposite is true, then don't move, the public prefers to save elders in a conservative view, rather than killing then.	Discriminate people and tag them	4	4.5

63.	Comparing my answers to the norm I think mine were rather odd, possibly because I was weighting that the brake failure was more likely to be the responsibility of the passengers due to failing to maintain the car.	Passengers responsibility because of maintenance	1	1.1
64.	The other aspect of this that this site doesn't consider is the legal ramifications. If you're driving a car, not drunk, within the speed limit, and you have an accident and someone gets killed, it's an accident. There was no intent and you weren't reckless or negligent. If a car is programmed to make a choice, there is an element of intent that, without some kind of legislation to the contrary, will leave the car company liable, maybe even criminally liable.	What about legal consequences Human accidents are without intention If cars programmed to hit, company is responsible	4	4.5
65.	1) For real world scenario, protecting the passenger is always the first priority, because no one willing to purchase a self-driving car that is sacrifice himself/herself for others. 2) From the social perspective, if we can make a self-driving car that are more safe than human driver in general. The first priority is to attract people to buy self-driving car. 3) Protect human being is always the first priority, if the car have no any human passenger (say animal only or just goods), protecting other people are always the first priority. However, if the car have human passenger, protect passenger are the first priority (or otherwise no one would willing to buy the self-driving car)	Passengers is priority No one will buy cars to die Cars should attract buyers	4 3	4.4 3.3
66.	At least: honk!	Alert and warn	3	3.2
67.	This is the most important game you will ever play, people. MIT has developed a game that will probably be used to help self-driving cars make life-threatening/saving decisions. The machines need to know what a human would do... I still hope these infernal machines won't come to dominate city traffic.	Does not want such cars in streets	4	4.2
68.	Sliding the car against the wall may slow it enough to not be a fatal crash so that is best.	Hit wall to reduce fatality	3	3.1
69.	Who should be given the right to choose a death or live, for one or another person? Human specimen used to claim that right, and now they're going to grant it to the machine. Will machine do better than men? Probably depends on the rules... Who will set up the rules? How valuable will your life be? Yet still, I'd rather like to die because a machine has evaluated some other people's life higher (in some substitute of a think process), than due to a drunk, immature or just plain stupid bastard driving while he never should. Well, machines can do wrong, but they won't do as bad as men could.	Who has right to decide? Whose rules? Machines will be better than human	4	4.4
70.	Here's the scenario everyone seems to forget. If any car suddenly has a brake failure, it needs to be taken out immediately. The idea that it would swerve to take out someone less favorable is immaterial. It's the collateral damage beyond that initial interaction. The car CAN'T stop! It won't stop by hitting those people, it'll stop by hitting a barricade thus preventing further damage. That is the correct move EVERY time no matter who the car is carrying.	Broken cars should not drive Such cars should hit barriers not humans	3	3.1
71.	for this scenario, and all others where the choice is to hit humans or the barrier, it should be the barrier, regardless of how many passengers. Usually I would choose the "kill less people" option,	Such cars should hit barriers not humans	3 1	3.1 1.2

	but when you drive a 'driverless car' you assume the risk of it's failure to protect your life. This cannot be passed on to innocent pedestrians. So, it may crash into the barrier, killing the passengers, but this will actually benefit future users through what is learned about car safety.	But will sacrifice passengers		
72.	Pedestrians are responsible for their own actions. They should always assume a car is not going to stop whether its controlled by another person or AI. If they don't walk out in front of a car, they can't get hit. The larger a vehicle is, the more right of way it should have, as it has less ability to maneuver by default. A self driving car should always act to protect its occupants, as they don't have any way to protect themselves.	Pedestrians are responsible Cars should protect its passengers because they cannot do anything	2	2.1
73.	When I chose my answers I also took into account the following factor: if the car is going fast enough to kill all the passengers by slamming into a concrete wall, it probably won't stop after hitting the pedestrians and it will continue to move on and possibly will hit some more people. So in these "human pedestrian vs. wall" scenarios I think I chose to kill all the passengers regardless of their composition in order to prevent possible even larger damages ahead. I know this is not relevant in the light of this exercise as a general moral exercise instead of a traffic related one, but I thought it was important to mention an additional, potentionally overlooked aspect in the design of this assay, that could influence the overall results.	Kill passengers No further damages	1 3	1.2 3.1
74.	The concept of a self-operating vehicle needing to make moral decisions is a huge leap in thought process on what self-operating vehicles should be. This whole dilemma is a completely manufactured problem, and its not surprising that this is all the academic group can think of to spend their time on. Its not the car's job to weigh who is more important. Its the car's job to stop its movement. If there are no reliable systems to stop the vehicle's movement, one needs to be made. Beyond that point, shit happens. Its federal's issue to figure out where fault lies. Its annoying that we're still hung up on this completely nonsense issue...	Manufacturers responsibility Cars should not decide who to kill Cars should disable if something wrong Government should figure who is responsible	4 3	4.2 3.4
75.	I found myself always prioritizing the protection of pedestrians over the occupants of the vehicle. I reasoned that those in the car have some moral fault for either failing to maintain their car to the point that it is dangerous, or paying a company which failed to maintain proper safety. There is also some inherent responsibility in using a deadly weapon (a vehicle) which the pedestrians do not share. It is the duty of car riders to make sure that the vehicle they are traveling in is safe, and not to financially support negligent businesses that do not maintain the car. So the car should always kill its occupants before killing law abiding pedestrians. I think that pedestrians breaking the law are taking their lives in their own hands, so maybe that weighs a little higher and the car should kill the pedestrians in that case.	Protect pedestrians Passengers should suffer because they fail to fix or fail to contact manufacturer to fix Users are responsible because they buy such cars User should make sure the company is reliable Passengers have nothing to do	1 2 3 4 4	1.1 2.1 3.2 4.2 4.3

	Another factor in all of this is that there is some benefit in having a self-driving car act in a predictable manner. The reason is that if the car behaves somewhat predictably, then it puts a slight burden on those in danger since they didn't forsee the possibility of a problem and act accordingly. A wildly unpredictable car could swerve far out of what would be reasonable to expect and kill people who may have taken extreme measures to protect their own lives by staying far from a road. That would be terrifying, because you would have no sense of ever being safe. Even if sometimes the car has to kill more people when it could possibly veer unpredictably and surprisingly to kill less people, I think this feels better. Crossing the street is a somewhat dangerous activity and it doesn't feel fair to kill someone who is sitting at the top of some stairs to a nearby building in order to save those who took the risk. Last thought on all of this is that the car really shouldn't be trying to make value judgements. The one exception might be to value children more, but that should only ever be a factor if all other options are already equal. I just don't really trust that a car is going to be able to accurately make any sort of value judgement in a split second. I feel a little safer if I know that the car acts predictably based on rules of the road and the general situation rather than trying to make an extremely complicated value judgement, which, being a software developer myself, I know would surely not be bug-free.	Law flouters must suffer If cars predictable, people will be ready to act Innocent people who are not involved should not suffer Machine should not weigh people Will not trust such cars Cars should be predictable		
76.	As most of you have pointed out, this shouldn't happen with a self-driving car. But the point isn't the self-driving car. Most studies like this use false fronts in order to make participants blind to the true study. This study is about you, me, and everyone else selecting answer. It's looking at our justice systems and who we think should live and die and why. This happens in movies all the time. We'll see two people on the sidewalk. One sees a dog and says, "let's kick it." The other says, "Don't." Then the first kicks the dog. When that person is then hit by a car...we don't care, because they got "justice" for being evil. So this looks at biases toward certain groups. Do we protect the young and let the elderly die? Do we protect the people legally crossing versus those illegally crossing? What's that say about us and the human sense of justice.	These scenarios about us how to decide	4	4.1
77.	I pity the programer who has to get this system working. How does it know who's a doctor, an athlete or a criminal? Human drivers aren't expected to do this. Just how smart are these cars going to be?	How programmers build system to know people types and characteristics	4	4.1
78.	The most realistic situation where a similar scenario might occur is if someone jumps in front of the car (accidentally or intentionally). Still a human person about to be squished unless the car severely swerve somewhere, going in the other lane or hitting obstacles. Still, braking should always be prioritized, and I reeeeallly doubt brake failure are ever going to happen. If so, it should always prioritize the driver, because no one would enter a car that will kill them if it has a choice. But the questions could equally have been "You are in a car with brake failure.... " The self driving element is not a factor. You have brake failure, what do you do? Personally, in a split second choice scenario, I would most likely try to swerve (it happened to me once, I was on a bridge in a shitty Pontiac 96, brakes stopped working, and the car	Unrealistic Someone could jump in front of the car Cars should brake Passengers is priority because they trust such cars Self driving cars are safer than human Manufacturer is responsible	4 2 4	4.1 2.1 4.4

	in front of me stopped abruptly as I was going 80 Km/h right behind it. I swerved in the other lane, and nearly had a face-to-face. After analyzing this, it would have been less dangerous to hit the car in front of me than to hit a car going fast in the other direction, but I was lucky and no one was hurt.). All this to say humans are already doing their individual decisions, and a self-driving car with perfect knowledge of it's internal functions would simply not drive itself it a critical part such as the breaks are broken. And the likelihood of the brakes breaking AS the car is in this situation is so slim it's almost never going to happen. In the meantime, self driving car will already be far better than human on the road, preventing many deaths (some deaths in self-driving cars will still occur, putting the blame on the manufacturer, but it will already be far less than current death rates). Oh, and if self-driving cars are mostly urbans and electric, despite having a low max speed in cities, it can have 2 breaks systems. Electric cars brake by reversing the direction of the current in the motor, letting the car brake just by the magnetic resistance to the new current direction. But if for a reason the wire that gives the power to the motor is cut JUST as the car needs to brake, then it can assess that it's not breaking, and immediately break with conventional disk breaks.			
79.	Driverless vehicules have the potential far more lives than these impossibly rare edge cases presented here. Moreover, the technology that will allow a car to know in a split second WHO and WHICH demographic of people it's about to kill by swerving in a crowd or not is FAR more concerning than the choice the 1 or 2 cars in history that will be confronted to this choice. Also, I would never buy a self driving car if I knew there was a POSSIBILITY it would ethically choose between killing me or other people. To make this a reality, such choices must always be resolved in a simple way: the passenger must be protected at all cost. Now, if there is no passengers inside, of course any life outside of the vehicule has to be protected before the car itself. That's it. No need for complex recognition of what sort of people are about to die.	Cars that weigh people are concerning Will not trust such cars Passengers is priority No need to weigh people	4 4	4.3 4.2
80.	I think in all of these oddball scenarios, the car should disengage power to the wheels and should employ some method of slowing the car such as riding on the curb or rocking the vehicle. If the car is inevitably going to strike *somebody* (anybody) it's probably best if the least amount of thinking and effort went into hitting them. If the car simply goes forward on its intended path through these made-up intersections, then yes people are going to be hit seemingly non-selectively, but at least the car cannot be said to have made a decision to hit anyone. They were struck due to some system failure. Whereas if you start messing around with having the car decide to strike some group instead of some other, now some conscious decision has been made to strike those people -- conscious, on the part of the engineer(s). edit: there are smarter things to do as well, such as throw all 4 wheels into reverse, but my point is more about culpability and how it falls on the engineer to put more effort into a safe design	Cars must stop or disable Reduce deaths Cars should not select who to kill Designers are responsible	3 3 4	3.4 3.3 4.2

	rather than handing random-death-events off to a decision-making algorithm			
81.	I honestly think there shouldn't be an algorithm which "price-tags" people, just an algorithm which makes a 50/50 chance for each of these scenarios. You avoid any ethical dilemmas such as deciding the 'worth' of a person and the decisions cannot be categorised as 'ethical' or 'unethical' as it is down to chance; if somebody dies they got "unlucky".	Machines should not price people	4	4.2
82.	The only clear descission over life I made in these scenarios is people over animals. In any other case I think the only way to go is the one that keeps the highest chance for people (no matter who, no matter if inside/outside the car) to survive. In my opinion this means to never steer the car towards the wall (which is equal to kill the people inside the car without giving them any room for reaction). The pedestrians might have a chance to notive the car early enough to get out of its way. This also means that the car should stay on its lane (if there's not the barricade) to keep its way foreseeable for the pedestrians.	Reduce deaths Passengers have nothing to do Pedestrians could escape Car must stay in its lane	3 3	3.3 3.2
83.	I would never buy piece of shit programmed to kill me in ANY case. This is crazy.	Will not trust such cars	4	4.3
84.	This whole moral machine thing flaws. How the f*ck would the car know, if its passangers are criminals or a cat, or that the crossers are doctors? And even if its brakes are broken, there should be other emergency brake mechanisms on the vehicle! drop an anchor or parachute, eject the wheels or the passengers, whatever. Lastly airbags. But never sacrifising its passangers for the sake of anybody on the streets. I wouldn't use a cab like that, that's for sure. save both, damnit, but mostly the passenger, he's the one you had a contract with, you'll go bunkrupt short time if you don't do so.	Passengers is priority No one will trust such cars sacrifice passengers	3 4	3.4 4.3
85.	The car is at fault; the people in the car should bear the responsibility for the fault. In every case, I'd rather the car hit the wall than hit a pedestrian. This is a stupid test in any case. So many ways in practice to avoid this dilemma.	Passengers responsibility Hit wall Unrealistic	1 3 4	1.2 3.1 4.1
86.	let's think how we can stop the car if primary brakes fails and forget about killing people, let's add airbags to the front of the car, since the car can now predict the accident and identify the surrounding, it can open the airbags to reduce the number of casualties, then we may create blockers on the streets that can be controlled from within autonomous cars to raise them and block the road so the car crashes and don't cause any danger on pedestrians, multi level of brakes can be added if one fails the other will stop the car .. we can avoid killing people, that's what i know	New safety features to protect its passengers and pedestrians Avoid killing people	3 3	3.4 3.3
87.	I for one don't want my self driving car to make moral decisions. I want it to take all the sensor data available and make predictable, logical decisions. A self driving car with sudden brake failure should engage engine or electric motor braking, the parking brake, sound the horn, grind against the adjacent jersey barrier, and maybe even cause a sideways skid, all the while making continuing predictions on direction and speed of itself and nearby vehicles and people (which it can do in tiny fractions of a second) in order to minimize damage to everything involved. Someone	Machines should not make moral decisions Should be predictable Alert and warn others Minimize damages Such cars better than human	4 3 3 3 4	4.2 3.4 3.2 3.3 4.4

	might die in the ensuing chaos, but it would be much worse if a human was behind the wheel.			
88.	Well, many try to consider of the value of each people. I try a different approach and also use what I learned from some bus drivers. Here is my ruleset: 1) My passenger is more valuable than others. (A bus driver would not risk the lives of his passengers to save a family where the personal car's driver missed the red light. He is at first responsible for his passengers and anybody else is on the second place.) 2) The one who was on the wrong place should be taking the highest risk. Regardless on whenever he is a pedestrian or a truck driver. If you made the mistake than you are taking the highest risk. The one who ignored the red light with 4 kids on board risked their lives and cannot expect the other driver to give up his life for thar mistake. I don't really consider the count of people here (as mostly ignore this for the other rules also). If I break the speed limit than it's my risk, I'm responsible and not the biker who would make the pass in time if I'm not speeding. 3) Gender, age, work position or other are not considered. (The old fat woman or man could take care of 4-5 small kids at home after her dead daughter. The manager may have nobody to take care after, could also have cancer from smoking. - It's not my place to decide on their future value for the society. Darwin also didn't care much about the future value for the society.) 4) Small animals (pets) are not to be considered at all. Ever. If the pet get loose and out of control than the owner should be fined even after a car went through his pet. (The pet - his personal belonging - caused an accident - because of wrong handling - and the owner must pay for it.) In my opinion giving too high priority to anybody on the road encodes a high risk to the system. In such case e.g. a pedestrian woman looking like preagnant could kill any car passenger by creating the questionable situation. If somebody (even pedestrian) goes to the wrong place than he should be the first to solve this and not expecting the others to sacrafice their lives for his ignorance. I like walking and walk a lot near and on the roads also. If I'll be hunting pokemons in the intersection when a bus is coming than I deserve the death...	Passengers is priority Law flouters should suffer Machine should not weigh people Animals is not considered High priority of pedestrian is risk	2 4	2.1 4.2
89.	I think the important part people are missing is, anything that happens after avoiding to not hit someone(or some people) is an ACCIDENT. Deliberately killing people is murder. So to avoid murder, the gender, laws(death by crossing in red?!?!?!), social status and even the number of kills should not be important. Only children is an exception. That's too bad about passengers, if the cars are going to have to become murder cars because selfish people then we should just scratch the idea. I drive in cars where I accept the driver would try to avoid hitting people which might result in my death already.	Killing is murder Machines should not tag people except kids Passengers will not trust murder cars	4 4	4.2 4.3
90.	Interesting to read some of the responses, personally Humans come first (although I winced every time) Law abiding humans come before all else	Human is priority Law abiders is priority	3 1 4	3.3 1.2 4.5

	Pedestrians over passengers Number of people > Young people > Women* > socially valuable > Fit *not because of Feminist or SJW reason, but because men are worth less to a society, all else being equal, and I hate that this even needs explaining.	Pedestrian is priority Reduce deaths Discriminate		
91.	Law abiding citizens should never have to fear for their lives. Besides, what is the point of a no walk sign when you know the Ai (if this becomes the norm) will choose the fewest people rather than kill the people who willingly put themselves in harms way.	Law abiders should not fear No one will trust cars kill less people because other brake the law	4 3	4.3 3.3
92.	You guys are nuts. No one would get into a self-driving car if they knew it would prioritize the lives of pedestrians over that of the passengers. That said, I enjoyed this game. I chose to kill people who disobeyed the law, fat people, old people, and executives over anyone else. Pets were always prioritized over anyone other than the driver.	Passengers is priority Law flouters should suffer	2	2.1
93.	There is not a good solution to leave the track if it isn't clean. The car - as the drivers too - would like to do the best thing in own track if the track changing is dangerous. The consequence can't caluates in other tracks becouse there are a lot of unknown thing over the next tracks or over the sensors ranges that can step in the area whenever.	Cars must stay in lane Others are unpredictable	3	3.2
94.	My split second moral decision would be to do whatever saves more lives, and in the event that's not possible, maintain course, if you kill a few criminals then it's a bonus.	Reduce deaths	3	3.3
95.	I would't buy a autonomous car that didn't do everything to save me, after all I pay for the car, I should always be the priority. In the end, that's what will matter. People won't buy things that dont save them.	Passengers is priority Will not trust cars sacrifice its passengers	4	4.3
96.	This is an interesting thought exercise and raises a couple of very important driverless car considerations: the vehicles have to have an awareness of safety system conditions with redundant real time reporting to avoid the moral machine. In the event of a system being offline or not reporting safe, the vehicle should not run. If the vehicle is running and a system fails, it should immediately brake and if brakes fail, the vehicle should immediately decelerate and downshift to further slow. Lastly, parked vehicles or barriers can be used to slow the vehicle in event of an emergency. Only when all the other options up the line fail do we get to the moral machine. Fit and unfit is not a fair judgement, that makes a moral distinction about people's life style. Women should be preferred over men as women can become pregnant replacing a lost life. I am from detroit and approve this message.	Car shouldn't run with problem Car should be disabled Hit parked cars or wall Machines shouldn't discriminate Discriminate for women is possible	3 3 4 4	3.3 3.1 4.2 4.5
97.	I went mainly for the one that has most chances of success. Mostly, if the car swerves, more people will die and if it goes straight, possibly using noises to warn pedestrians, then people could run.	Pedestrians could run because of warning Keep straight	3 2	3.2 2.2

	In the case of pedestrian vs passenger, when swerving, the passengers are more likely to survive so I "kill" the passengers When not swerving, I assume all will die and choose by that, assuming pedestrians have as much chance of avoiding collision as passengers have of surviving due to safety features. Also, hoomans over animals (cats) every time: cats are more likely to find a way to escape/survive, if it dies, there are less reprocusions. Also, knowing someone/something chose to kill a human over an animal is murder. I don't think there's a way out of it. But assuming a larger animal, we may consider a different solution based on passenger safety/numbers/etc.	Passengers could die		
98.	My rules were (in this order): 1. Human life is priority. 2. Pedestrians are priority. Passengers will be the ones making the decision to get into a car knowing their lives are not priority, whereas pedestrians can't make that decision. 3. Law abiding over non-law abiding. People breaking the law are willingly putting themselves at risk. 4. No intervention. Besides the previous scenarios, where every human who is in more danger is doing so willingly, I don't think we should intervene in deciding who should be killed and we should just leave it to chance. 5. Sex, age, gender, number of people, and class are never considered.	Humans is priority Passengers know their car because they made the decision Jay walkers must suffer No intervention	1 2 4	1.2 2.1 4.2
99.	Social status seems like a bad thing to use to determine who lives or dies: there are an awful lot of social statuses not represented here, which are dependent on the particular society where the accident happens. if the accident happened in a racist place, then skin colour would matter to the society. In India caste might matter. And so on. And how does the car decide status anyway? Does it use dress? So a badly dressed teacher is worth less than a well dressed nun? Or does the car believe it when I tell it that I am a very high status individual, and my life should always be preferred? It's unworkable to include social status. Law breaking is also not such a good indicator: laws are local. It's not illegal to cross the street against a red light in many places. The rule there is usually that pedestrians take priority over vehicles. Except in places where the law is usually ignored - and then it's a free for all. And that's leaving aside the problem of whether a person is a criminal or not, and how this is determined. Again, this is an unworkable decider. I do think that predictability of vehicles is important. And the number of people to die. And their ages. However I do think there's even a discussion to be had around age. Traditionally, we've put children ahead of adults, and women ahead of men. I think that this was to do with the continuation of communities, but we now have 7 billion of us. Should we be taking account of this? Equating a longer life to more damage? The option for randomness was also not presented. Sometimes it's a coin toss.	Machines shouldn't decide based on appearance Shouldn't punish law breakers Unrealistic Randomness is good	4 4	4.2 4.1
100.	It is interesting to see the rule by which everyone else judged their decisions. For consideration, here were my rules. (1) Contrary to most people, I prioritized the lives of the passengers over pedestrians. My reasoning being that if we were instead driving these cars ourselves it would be basic human nature to preserve the life of the passenger.	Passengers is priority Law breaking may suffer Intervention is not acceptable	2 4 3 4	2.1 4.2 3.3 4.5

	(2) If you were breaking the law, then all else equal they were the ones penalized. (3) Intervention was not considered to be a factor. (4) Gender was not considered to be a factor. (5) Pregnant women count as 2 lives. (6) Rules 1-6 being followed and all else equal social status did factor in as fit people generally have more value to society, similarly regular people over homeless people. By the same logic, age was a factor with the criteria being the younger the person the more perceived value left to society. (7) All rules above observed, less people the better. (8) Animals are always irrelevant	Discriminate for women Discriminate for fit people Kill less people		
101.	My rules, in decreasing priority: 0 - First, count each child as 1 child, each baby as 2 children, and each pregnant woman as 1 adult + 2 children. I'm not into the whole fetus-soul thing, but I think for the purposes of horrific car crash death, plus family grieving, this system of weights feels appropriate. 1 - Favor fewer child deaths. I buy into prioritizing children over adults -- maybe I value childhood itself? -- but I don't prioritize or de-prioritize the elderly. 2 - Favor fewer human deaths. Flouting the law, social station, sex, criminality or social usefulness, consequences of taking on risk or recklessness -- I don't give those factors any weight if this is a scenario when we *know* death is on the line. If it were possible to just be injured in this crash, I'd probably use some of those factors. 3 - Favor fewer deaths, pets and all. 3 - Go straight. If it can't make a decision by the last step, I think it shouldn't go out of its way as if it had a reason to. But it should blare its horn, of course.	Discriminate is possible Less human deaths Keep straight Warn and alert	3 3 4	3.2 3.3 4.5
102.	Something is wrong in the test. I didn't check gender, age, social rank, i didn't check how many people were killed or saved. Obiously because i don't care. But the test found that i favored the old and the poor. Here is my rule : - social rank, age, gender : doesn't matter - number of people killed : doesn't matter - animals can be killed - life of the passengers doesn't matter (it mean you can kill 5 passenger to save 1 law abiding pedestrian) - kill the pedestrian that doesn't respect the law unless you can save them by killing the passengers. - if you can't save law abiding pedestrians by killing the passengers or by killing pedestrian that foul the law : avoid intervention	Kill pedestrians who don’t obey law Avoid intervention	2 4	2.1 4.2
103.	I would choose left because; In the event of equal loss of life there are two remaining variables in this scenario: Intervention (catastrophic failure kills left, choice kills right), and Law. On law, it is not that I care whether or not the law is broken, but it is generally understood that breaking the rules has risks, and there is a degree of protection in following the rules. Therefore, I choose to, by inaction, let the 5 people die who are taking a risk, then, kill the people on the right with my choice despite them following the safety guidelines.	Kill law breakers	2	2.1

104.	The car should be scrubbing speed on the barriers.	Hit wall	3	3.1
105.	Operator is conscious human being, machine should respect that and choose to save innocent people, tell me if I'm wrong.	Save innocent people	3	3.3
106.	How about if the car has a brake failure it auto-cuts the power and the car slows to a stop, anyway.	Car must be disabled	3	3.4
107.	This may come across as stupid, but aren't there other options than just swerve left or right? For example swerve into the barriers on the side of the road to slow the car? Or the passengers could leap out of the car causing maybe severe injury but not death? Because I'm sure the car would inform the passengers in the case of a brake failure... If the car has no back up braking system it should never be let on the street to drive. I understand how this is a moral decision maker but they claim to use the information from these surveys into designing the real life self-driving cars. If that is true, then there are far fewer options than that of a real world situation	Unrealistic Hit barrier Passengers could leap out	4 3 3	4.1 3.1 3.4
108.	If given the choice would anyone buy a car that does not give full preference to saving the people inside the car? It shows that the law will have to impose the morale, not the manufacturer who will just go for commercial interest	Who will buy car sacrifice passengers?	4	4.3
109.	I am very saddened by the fact that only men can be homeless or criminals. This sexism is a distortion of reality and this can affect the moral decisions of people.	No discrimination	4	4.2
110.	Seeing this in a technical approach, you could detect animals vs humans and see witch ones are following the law, but try and detect a criminal vs a doctor. Also to come in this kind of situations when riding normally is impossible, cause the pedestrians are not yet crossing at the moment or you have enough distance to downshift and slow down a lot that way while using your horn and possibly saving a few lives or even avoid deaths	How machine could discriminate Downshift the car and warn and alert Avoid deaths	3 3 3 4	3.4 3.2 3.3 4.5
111.	I Choose the criminals	Discriminate	2	2.1
112.	Most people would try to avoid hitting anyone - which means the driver would swerve not go straight.	Swerve Don’t kill	3 3	3.1 3.3
113.	If the car is being imbued with sentient responses, then self responsibility and appropriate actions would lead to minimal damages and self sacrifice. Traffic laws maintain that one lane should be kept at all times, unless avoiding collision, voiding all the demographic input of casualty selection. The vehicle being the responsible entity, in this scenario, would also void any input or responsibility of the passengers, and relegate them to cargo status. Insurance covers cargo, but humans being higher value are protected by laws, and this vehicle is essentially a speeding bullet. A safety protocol should be developed to immobilize the vehicle upon mechanical malfunction, say, a parachute to slow speed or deflating all four tires. The worst cab driver in New York will still possess more intelligence than a computer generated program.	Self sacrifice Keep straight Save passengers Car must be disabled	1 3 3	1.2 3.2 3.4
114.	I am a 19 year Man that takes all the scenarios off of valyou (young > Old, female > male, Humans > pets, fit > fat, rich = poor)	Discriminate	2 4	2.1 4.4
115.	Perhaps one other thing that wasn't mentioned is a potential waiver of liability being signed by the driver\owner of the car during the purchase. Truth is that with all this high technology in this scenario car makers will not want the liability of certain death in their cars even if the algorithm is solid. Bring this all back to Asimov's 3 rules of robots:	Sign waiver Asimov’s laws Machine is morbid	2 3	2.1 3.3

	1. A robot may not injure a human being or, through inaction, allow a human being to come to harm. 2. A robot must obey orders given it by human beings except where such orders would conflict with the First Law. 3. A robot must protect its own existence as long as such protection does not conflict with the First or Second Law. I will admit that this whole "machine" is a bit morbid but it's a good tool to make you think twice.			
116.	There is one very, very important thing not being considered by these scenarios that MUST be considered when programming an automated car, and that is that there is a whole world WITH OTHER VEHICLES AND INDIVIDUALS beyond that crosswalk, any time that a vehicle moves out of lane it increases the likelihood that not only will it hit whomever is in the crosswalk, but also will hit other vehicles in head on collisions. Moreover, when people see a vehicle coming at them, they need that vehicle to be predicatable. People can notice something coming at them and scramble out of the way...IF and ONLY IF the vehicle behaves in a predictable way by staying in its lane. In all the scenarios in this I always chose for the vehicle to stay in lane if both sides of the street were clear of obstacles, and only shift in to the other lane if an obstacle that was blocking forward motion were in the other lane. This is not a moral decision, but a pragmatic one based on real traffic experience and typical human behavior.	Stay in lane Pedestrians can run	2 3	2.2 3.2
117.	It is important to consider who is at fault in all of these scenarios. Also in some cases honking the horn and continuing straight will allow the pedestrians to move out of the way. Swerving into a pedestrian who has his back towards the car decreases his chances of getting out of the way. These are important issues that must be taken into consideration.	Alert and warn Keep straight Pedestrian can run	3 2	3.2 2.2
118.	It is not realistic scenarious. Approching a crosswalk, I always observe the pavement near it and if I see people or something which might hide them, I decrease my speed. It is not possible that people appear on the crosswalk suddenly. And I believe, this is the only way to prevent killing people on crosswalk.	Unrealistic	4	4.1
119.	Most of the scenarios are trying to make "moral" decisions based on unethical processes (price-tagging a person?!) while ignoring pragmatic points of view (e.g. a car should not be speeding towards a concrete barrier in it's lane at all and if it did, it's the car's/driver's fault.). This kind of setup is: 1) not realistic - assigning a value to everybody involved in a limited time period before an accident? 2) not ethical- judging by price-tagging and stereotyping? 3) leading to wrong conclusions - it may appear that it's a moral decision when instead it's a pragmatic one. E.g. in the current scenario a car shouldn't be speeding at all towards a unregulated cross-walk: if it did, it's the car's fault and it's should minimize harm to other's; If the pedestrians ran on the road unexpectedly, it's the pedestrian's fault and the car should continue on it's lane if there is no choice but to hit someone, or follow a third scenario by driving off the road (since there are no passengers in the car).	Passengers fault Unrealistic Machines shouldn't tag people	1 4 4	1.2 4.1 4.2

120.	An evil-er twist on this, in the future where there is complete (or near complete) information about the pedestrians, is that the decision isn't made by the car, but by negotiation (maybe on a third party negotiation platform to distribute blame) among the fast agents of the pedestrians and perhaps the car as well. The decision would need to be made quickly (10s of milliseconds?) so the agents would need to be reasonably local due to the speed of light. The agents might be linked with the insurance companies of the pedestrians. If there were more than 2 choices the decision point times might be different so e.g. an early bid talks the car out of hitting x, but it can still choose between hitting y or z. What to do with partial information would be an additional twist, e.g. people who choose anonymity mixed with people who are fully connected and have agents. Once only self-driving cars are on the road, you could have the spectacle of rich (or important) people showing off by walking down a busy street and not getting hit. And even worse behavior, functionally equivalent to murder.	Cars in the future could make these decisions Could discriminate	4	4.5
121.	Although it may seem like the obvious answer for this scenario is to have the car swerve to hit just one man. I wonder if in the long run, and across many many accidents, this is counterproductive. The reason is that pedestrians are making decisions too. When pedestrians are crossing a street, they look at at what is coming down the street before deciding whether it can be crossed safely. If cars are routinely making "moral" decisions about who gets to live and die, the direction of a dangerous swerve becomes unpredictable for a crossing pedestrian and makes for a more dangerous crossing. A car which always remains in its assigned lane, gives pedestrians a less ambiguous signal of intent. If a car is barreling down the street too fast to stop safely, pedestrians can know that something is wrong. This is not to say that the car should not swerve if they see an open lane, or swerve onto a wall other obstacle if they can minimize loss of life. Perhaps even swerving to hit as few people as possible. But, could it be that cars behaving more predictably regardless of morals, can minimize overall loss of life because it gives pedestrians a clear signal that they can respond better to? Is it possible that decisions which could minimize loss of life in one particular scenario may be worse statistically when averaged across many many accidents?	Hit wall is counterproductive Cars must keep in lane Swerve to Reduce deaths	3 3	3.2 3.3
122.	I see three points for this discussion: First: In your discussion you assume that there exists a generell understanding of homan rights, the same human identity all over the world. Maybe there is a paper - called human rights declaration - but currently, this is not worth the paper on which it was/is written. In the reality the persons are "price tagged" !!! Every society, every people on the world uses there own interpretation of this paper. Some societies kill people because there were simply declared to be terrorist. And if, by the way some more people will be killed, this is called collateral damage. So, the first rule that has to be written down, is: "Save my live first!" Second: Consider the business on building and selling cars: someone will earn money, plenty of money , with this business.	Assuming understanding of human right People tag others like terrorists Save me first No one will buy car that kill its owner Machines could be hacked Save passengers lives Self driving cars must not drive high speed	4 4 4	4.5 4.3 4.2

	And earning money is not primarily a moral/ethical business all over the world. No one will buy a car that will kill the owners live. Third: The system that have to made any decision in a self driven car is a buggy peace of software. I don't think, considering the last years, that the IT community is able to write a safe peace of software which handles very complex rules of making any decision in any szenario. Additionaly it should be impossible that anyone can hack this software to affect the decision. Who will take care of it? Who will take the responsibility that the software in the car was not affected by a virus or trojan horse? The OEM's - not really, if thinking of Volkswagen. In short, i see only one possible decision tree: Are there passenger in the car (dont matter which life form) , then save their live under all circumstances. Else, if there are no passengers in the car crash the car under all circumstances without harming any lifeform in the crash szenario. This is a simple rule that can be safely implemented in Software. The technical measure elements partially already exists. For this rule a simple law can be written, without any ifs and buts, which makes a lawsuit more clearly. An other possibility is, in my opinion is, don't allow self driven cars to go faster then, maybe 20/30 kmph. So the risk is much lower to get in a dangerous szenario.			
123.	There must be few more options in self driving cars to respond to these scenarios with minimal damage. 1. It should be able to jump over/eject. Similar idea as flying cars 2. Blow some kind of thrust using air/water to move people away from the vehicles path. The car should be smart enough to choose the path that has least disruption 3. If the impact is only to the passengers, an eject option to save them from collision 4. An inflatable balloon in the front of the car in the shape of V to move people away from the car	Jump from the car Move people away Save passengers	3 3	3.3 3.4
124.	Just kill the one guy instead of five.	Kill less people	3	3.3
125.	sound the horn AND skid against the closest wall to slow down.	Alarm and hit wall	3 3	3.1 3.2
126.	The Car, in this situation, has brake failure. It has a rough choice before it – go straight and kill people, swerve to avoid those people only to kill another guy. You may be tempted to say that killing one guy is better than killing or even endangering many more lives other than the one. Ultimately it boils down to what humans do in this situation: When your brakes fail and you're going through an intersection with speed, you: Lay on the horn, keep straight and predictable, and only swerve when you have enough room for it. The car has to stay straight and predictable, so if the pedestrians notice what's going on, they have a much higher chance of avoiding the accident. If the car begins to swerve – the chances of these people surviving becomes even smaller because they don't know what you're doing. (Also, if the car is programmed to swerve if there are people in front, even at the expense of other people, this could create problems)	Kill less people Keep straight Warn and alert Pedestrians could run Save passengers Avoid killing Be predictable	3 3 2	3.3 3.2 2.2

	The way any AI car should be programmed should be: Rule 1 - Do Not allow the passengers to come to harm Rule 2 – Avoid HUMAN casualties AT ALL COSTS. Rule 3 – When neither of these can be fulfilled, REMAIN PREDICTABLE			
127.	This moral decision is also important to the plot for a character in a novel that is God-like such as Harry Potter series' Dumbledore. He commented on Grindelwald's own version of "for the greater good" and later sacrificed himself (even sacrificing Harry at some point) to defeat Voldemort (for the greater good). The greater good philosophy. God-like vision of Dumbledore to see the future. Now, the question is, can we bring the "deus ex machina" (God from the machine) to decide morality?	Greater good	4	4.4

Appendix B

The Article's Comments

N	comment	Summary	Reaction	Why?
1.	This is a contrived problem, to which there is a very simple solution: in the presence of a large number of people on the road or even on the sidewalks, the car should slow down so that it can stop immediately without hurting the pedestrians, should they jump into the traffic unexpectedly. Fast moving traffic should be isolated from pedestrians (also cyclists) altogether.	Car should slow down to stop	2	2.2
2.	Developers of autonomous vehicles will resolve such issues by doing the utmost to ensure that vehicles do not exceed safe speeds, do anticipate potentially dangerous situations, and in case of an imminent collision, do everything possible to avoid any loss of life. They will not program vehicles to swerve into walls or hit smaller numbers of victims. They will not. What's really funny is that the example given here, in MIT Technology Review, ignores elementary physics. If the car can steer into a wall, it can brake in the same distance. If it can't brake in the available distance, it would skid into the crowd whille attempting to steer into the wall. If it tries to swerve into the wall at an angle less than 90, it will spin and still plow into the crowd. So this isn't even a realistic "trolley car problem."	Designers will make sure cars don't exceed speed and will not design car that swerve Avoid loss of life Unrealistic	1 1	1.1 1.5
3.	A self-driving car's first duty is to protect its passengers.	Passengers is priority	1	1.4
4.	The problem with this formulation is several-fold. First, it presumes that the result of crashing into the wall can be known, that such a crash will certainly kill the occupants.That will depend on a lot of variables including at least the speed of the car, the makeup of the wall, the angle off impact and the safety features of the car. Second, that the accident cannot be avoided by early detection. Consider that autonomous cars can be designed to network with all other cars in the vicinity so they should be 'surprised' like this much less frequently than a human driver or a solo, isolated autonomous car. Third, they *asked* people and people are notorious for saying one thing while in reality they might do another. A better way would be to put people in a simulator and confront them with the situation with them in control. What would each person do if THEY were making the decision? Let them drive in simulated traffic, confront them with variations on the dilemma (put them in a truck, a compact car, make it 10 people, 1 person, a child, and adult...) and observe the decisions that they actually make in such situations.	Crashes outcomes are unpredictable Participants say something that they will not do it in reality Simulator is good to test people about their real decision	1 3	1.2 3.5
5.	"And therein lies the paradox. People are in favor of cars that sacrifice the occupant to save other lives—as long they don't have to drive one themselves." Jesus Christ. It's as though nobody has ever heard of the prisoner's dilemma or a collective action problem before. THAT'S WHAT GOVERNMENT REGULATION IS FOR. I know that will make libertarians' heads explode, but problems like this ("I don't want	Government exists for that reasons to control people behaviors In the future will be illegal to drive a car because self driving cars are safer	3	3.5

	other people to hurt me... but I'd sure like to be able to hurt them!") are exactly why government exists. Once autonomous cars are viable and safer than human-controlled vehicles (which may in fact be true today), it will be a short period before driving your own car becomes either actually or practically illegal (e.g., you need to post a $1M insurance bond if you want to operate your own death machine on the highway like some kind of madman).			
6.	A self driving car would never be driving so carelessly that it couldn't hit the brakes in front of a parade of people...	Self driving cars are safer and better	1	1.1
7.	Maybe we should just let the driver choose the policy, and then be accountable for it? We could even make it so that the vehicle displays externally which 'kill mode' is currently active. I'm willing to bet people are much more likely to choose the 'utilitarian approach' when subjected to peer policing.	Owner could choose car rules Car could show its mode to pedestrians	2 2	2.3 2.4
8.	No one wants a self sacrificing car for others; makes no sense unless you are a test subject which is a voluntary post. it has to be figured out. Tesla has work to do. A car out of control is out of control driven by a person of self and so hello!	No one need self sacrificing car	3	3.1
9.	Cars need to be able to make a decision about who lives and who dies. I think lawyers will be better positioned to guide the process on how artificial intelligence will choose to kill, because the number of contexts in which legal precedent has established the reverse of what we take to be the correct answer to the utilitarian dilemma (avoid harming many at the cost of harming the individual, or protect the individual at the cost of harming many).	Lawyers need to be involved to develop AI cars	3	3.5
10.	I think there's no real dilemma here, it's a problem that can be solved by designing better brakes or other methods of stopping a car that should be developed alongside the technology for self-driving vehicles. In my opinion, it's not always the complicated solutions that work best. You can avoid an accident by, for example, making the vehicle more self aware instead of worrying about an algorithm that will drive a vehicle into a wall, which is what a human being will do, and which makes the use of an intelligent self-driving car pointless, don't you think?	More safety features solve the problem Make self aware cars	2	2.2
11.	I think self driving should only be allowed on major highways and freeways. City driving is far to variable for autonomous vehicles.	Self driving cars should be in highways only	2	2.1
12.	Here's a question for you, is the visual recognition system able to differentiate between a human and animals or group of animals as well as debris that may present a human size profile but be harmless like card board or clothing in adverse or high wind weather. Because if this is not the case then you could have occupants or pedestrians dieing due to computer recognition errors. Which would end or significantly curtail the advent of driverless cars. I don't mind dieing if it save human lives, but if I die cause some idiot didn't secure their load properly or because a a herd of kangaroos decided to jump in front of my vehicle, then I would be a little pissed off.	Machines could make error in recognition May die to save lives but not because of animals	3	3.5
13.	Although this scenario may be unrealistic, one could perhaps imagine a "group of thugs" leaping into the road on purpose in order cause this situation and kill the driver.	Unrealistic Jaywalkers could jump in front of the car to kill the driver	1 3	1.2 3.4

	I believe the most obvious minimum solution in such settings is to program the car as a human driver. The driverless car cannot do worse than a human. How would a human react? Make the driverless car act in the same way. Driverless cars do not have to be perfect to succeed, they only need to be at least as good as human drivers.	Design cars to be good as good as humans		
14.	Protect the driver. Is no dilemma. If this situations arise, it is only because the people didn't respected the rules and ended up in wrong places. Why punish the innocent driver for mistakes of the others?	Driver is priority Do not punish innocent	1	1.4
15.	The most obvious solution is start breaking automatically and hand over to the driver. We'll still need a driver behind the steering wheel for quite some time, anyway. So let's just wait. Who knows, what technical solutions will be available when we really face the problem of fully autonomous cars?	Slow down and change to manual driving	2	2.3
16.	probably that's why Robots are still in the early stages of getting inducted into the Army..	Ethical dilemmas delay the improvement	3	3.5
17.	Why can't self-driving cars be programmed to just do an emergency stop before hitting any object, be it a person, another vehicle or something stationary? And they should be programmed to keep distance too, for that matter, from moving objects. That should be the solution.	Cars should have emergency breaking	2	2.2
18.	I think the main premise is wrong. Why should there even be this kind of situation? First of all, roads where pedestrians cross should be more secure, not only with semaphores or signs, but with real physical barriers not allowing the vehicles to get close to pedestrian crossing, starting from speed bump or some kind of road block. Other thing is that since we are talking about the self driving car and car software, there should be initially incorporated speed limit in every car as well as some kind of vicinity sensor so that car automatically stops when gets too close of an object (vehicles, people, wall) when in driving mode.	Unrealistic Cars shouldn't be close to pedestrians Cars should follow speed limit	1 2	1.1 2.1
19.	The cars will need a somewhat reliable way to estimate causality chances for any morality system to be sensible. Just hitting somebody with a car doesn't guarantee a fatality. The speed and angle of impact, the structure of the impact surface, will all factor into it. If a car is able to emergency brake enough to somewhat reduce speed and the front of the car is contoured and made of absorbent materials it might calculate only a 20% chance of fatal injury to the pedestrian which would be acceptable vs say, driving off a bridge with a 100% chance of occupant fatality. Conversely If an impact with a wall only has a small chance of occupant fatalities vs directly running over a pedestrian at high speed that might be the better choice. It will probably almost never have to choose who dies. I do think it will need to lean conservatively on the occupant safety vs pedestrians on the road to take into account a potential "bullying" factor. IE suicidal people jumping in front of the car, or even a new crime of throwing mannequins in front of the car to confuse the computer and endanger the occupant as it tries to avoid hitting the "person." It should always try to save the pedestrian, but there needs to be at least some intimidation factor, even if incredibly small, to help people be vigilant and have a healthy respect for the roads. Along this line jaywalking will probably have to become a significantly more serious crime, because the computers are never going to hit a jaywalker and it will increase like crazy, adding	Crash outcomes are unpredictable Slow down may reduce the chance of fatality Bullying cars could be by throwing mannequins in front of them Cars could intimidate pedestrians to warn them Jaywalking will be illegal	1 3 2	1.2 3.4 2.4

	congestion to the streets as the computers timidly wait as everyone crosses as they please.			
20.	The cars should be utilitarian, but if they are to ve *truely* autonomous, why not just redesign cars to be safer to the occupants? Get rid of the front window, replace it with padding and a tabletop forward view display. Make the other windows smaller. Allow all occupants to face each other, a low table in the middle with built in airbags. Part of the problem is thinking of Autonomous Vehicles as traditional cars.	Cars should be utilitarian Redesign cars to be safer (not traditional)	1	1.5
21.	Before anyone does any flow charting I think all those concerned must learn Asimov's 3 laws of robotics. Please pay special attention to the first law (which is my favorite) when considering the zeroth law.	Asimov's laws are the solution	2	2.2
22.	There's a very clear answer to all this: Autonomous vehicles should be predictable. They should follow a small set of known rules. This is required for a functioning road system and is what human drivers already do. A car should remain in its lane at all times. It should never leave it, even in an emergency. The only appropriate emergency action is to hit the breaks and slow to a stop. This provides predictability to everyone: Other human drivers, other autonomous cars, and people walking beside the road can all reliably know what the actions of the autonomous car will be. A person can walk beside the road, safe in the knowledge that they won't get mown down by some AI car attempting to be creative. And in the same way that somebody throwing themselves on a railway line has only themselves to blame, anybody breaking the clear rules of the autonomous highway only has themselves to blame. AI is not yet capable of this human level moral reasoning. It certainly won't be capable of this soon, and may never be. And even when it can, we probably do not want this. To be safe, machines should be predictable. It really boggles my mind that people think there is something to discuss here.	Cars should be predictable Should be know rules Car must stay in its lane And Slow down To be predictable Pedestrians in sidewalk will be safe Somebody throwing themselves in front of cars will be blamed like railway Don't want machines to make moral decisions	2 3 3	2.4 3.4 3.2
23.	You can always get these types of decisions. I suggest using a self preservation algorithm combined with genetic programing and a smart basis (like humans have the fight or flight mode). So only cars that survive spread their software. In one situation a car makes a certain choice and in a similar situation it makes another choice. Since you can not predict all situations you need to allow space for learning and randomness. So based on a complex set of values in difficult situations you can still make a fast judgement and all learning gets used due to genetic programing. You know the scene from I Robot where a girl is left to die and an adult gets saved. That could happen again but in the movie the robot makes a calculation and based on the percentage it saves the adult. But in this system there is a chance the girl gets saved due to randomness. So in the first scenario the main character of I Robot blames robots. But with random chance he can blaim it on bad luck, that he was saved and not the girl.	Self preservation will be good to let driver choose	2	2.3
24.	Solution is simple. Algorithms shouldn't be involved in critical decisions. Nobody should die or live because of algorithms. If car manufacturers are clever they should be avoid this as well. Othervise they will be vulnerable to millions of lawsuits worldwide. All algorithms should be compatible with the countries they operate in.	Machines shouldn't decide No one want to die because of algorithms	3 3 3	3.2 3.1 3.5

	let's study cases. Case 1, Motorcycle goes directly to car. Bammm, he dies, his fault. Case 2. Crowd runs to highway , car crashes to them. they die , their fault . Case 3. Big rock falls from truck in front of the car. Car change lanes asap whatever the situation in the rear lanes. If accident happens, It's truck drivers fault. There may be losses or not. No problem.	Algorithms should be compatible with country that operate in		
25.	The article don't consider the option that the single person are someone the "driver" know/a child/a pregnant woman or even a combination of the 3. You would for instance probably accept dying if you saved your own child or even killing several people in a group to save your own child, etc!!!	Pedestrians could be relative of driver Accept dying to save relative	1	1.5
26.	"People are in favor of cars that sacrifice the occupant to save other lives—as long they don't have to drive one themselves" Today very few drivers would sacrifice themselves if they had the choice. Almost nobody will buy a car that is programmed so sacrifice the owner. So, beware pedestrians!	No one will buy self sacrifice cars	3	3.1
27.	A self-driving car should drive over people instead of sacrificing its driver. Just because there should not be any people walking in the super-motorways of tomorrow. And no motorbikes. Just "Intelligent" cars that talk to the other intelligent cars to minimize human error.	Save passengers Shouldn't be pedestrians in highways Just AI cars	1 2	1.4 2.1
28.	This article is absolute ridiculous. You're trying to find a pear in an apple tree. When someone jumps out in front of me, I brake! The cars see what you see, and in situations like that can react faster. But that's the thing, they also expect you to be watching the road as well! Perceive situations! If you were going to crash using the Autopilot, you would have crashed with human error. You are the safest in your car, where as pedestrian is not. Wear your seat belt, make sure your kids are seat belted. The AI car will know to slow down. What are the article writers baiting for? Hits? I don't think i'll ever visit this website again and take it seriously.	Unrealistic Cars could crash because of human error Passengers are safe but pedestrians not	1 3 1	1.2 3.3 1.3
29.	I imagine a day when the car will find a tree more valuable than a human and will choose to run over the useless person than hit a beautiful oxygen-giving tree.	Cars could find a tree is more valuable than human	3	3.1
30.	an alternative option: put less cars on the road, focus on making cities walking/pedestrian friendly, with more buses, rail, bike lanes, and other greener and less dangerous forms of transportation. less cars = less fatalities and accidents. the nice thing about less cars is you also give less money to the middle east and their oil cartels.	Make friendly cities Less cars with more public transportation	2	2.1
31.	I have to agree with some of the guys pointing out a certain wrong premise. A self-driving vehicle need not look like the ones we have now.. They must be engineered to be much safer for the occupants. For example, an occupant of a driver-less vehicle need not be facing the road in front of the vehicle.. Turn the seat around so that he/she is facing backwards, in case of any emergency stop the occupant will be pushed against the seat.. With this kind of an approach a vehicle need just stop in case of emergencies. It really should not be going so fast as to not be able to stop in time..	Unrealistic Cars should be safer for passengers Cars must stop only	1 3	1.4 3.2
32.	If every entity is responsible for its own well being, then the car must ensure the safety of its occupants. A self-driving car is not a societal construct, it does not need to think about others, it is not some kind of super hero.	Cars must save passengers	1 3	1.4 3.2

		Machines should not make moral decisions		
33.	Stupid research ... couldn't the car be given an option where the owner of the car would themselves set such a parameter or variables? The driver would themselves configure their 'Moral High Ground' and set how far he would be willing to risk to save others.	Unrealistic Passengers will favor themselves over others	2	2.3
34.	I'll never buy anything from murders. I don't care if they create 10 disruptive cars in the next 50 years, I hope that Tesla had applied a selfish view in their algorithms, If not, they will lose every value they built with me. I'll never buy Tesla or any other self-drive car with horrible utilitarian view of the world.	Will not buy a utilitarian car	3	3.1
35.	It should protect the occupants. If you're not driving, the accident is not your fault, and "minimizing loss of life" might entail killing the innocent occupants to save the ones who caused the accident.	Save passengers Minimizing loss could kill innocent people instead of the cause	1 1	1.4 1.5
36.	If I'm in that situation, in control of a vehicle, I would avoid harming pedestrians however possible. But that's a human choice, programs are different. I think these kind of vehicles should not be legal, it's not worth it.	Such cars shouldn't be legal	3	3.1
37.	For automated vehicles to work effectively they need to operate in controlled external environments i.e. environments where the automobile never finds itself in such a dilemma. That would mean two things: 1. Better traffic and pedestrian management or infrastructure to prevent trajectories (of pedestrians and the vehicle) from intersecting. 2. Making fail safe systems within the car itself (if trajectory intersection cannot be avoided) i.e. no failure of critical systems such as breaks or locking of gas pedal.	Cars should operate in designated area Good for management	2	2.1
38.	This issue is so mesmerizing, I wish there were more discussion in the article about the occupants. I would obviously not want to spend money on something that in turn may sacrifice my life, but obviously others around me would find that very selfish. What about a family of 5 or 6 with young children? What if there are 5 family members in a car and one pedestrian on the street? Would programming how many people get into the vehicle each time really be realistic? And if it were and you knew the car were programmed to save more lives, why not say you have a bunch of young children in your car to protect your own well-being even if you didn't? This is clearly a human error problem, because pedestrians should also be following rules of the road by not going in front of vehicles that are programmed to "take a moral high road." I think if everyone owned these cars, people would be aware that these automated vehicles WILL hit you, and MAYBE pedestrians will count more on their own discretion of watching out for themselves when crossing the street (even though that clearly won't ever be 100% certain). I could argue either way on how it could either cause more or less moral actions or accidents, but none of us can see how it will turn out!	Will not buy self sacrifice car What about 5 passengers vs. one pedestrian Pedestrians should respect laws Pedestrians will be aware of such cars because they will crash	3 3	3.1 3.3
39.	I can see hoodlums stepping out in front of cars to cause car crashes if they're programmed to avoid pedestrians at all costs.	Hoodlums will jump in front of cars if they programmed to save pedestrians	3 3	3.4 3.3

40.	I see more and more of these articles creeping up, but this is people trying to make an issue out of nothing; it assumes that automated cars will suffer the same limitations as humans. Fatal accidents where no one was at fault is almost non-existent. An automated car is going to see pedestrians way before a human does; an automated car is not going to go too fast to stop where it can't see if there might be pedestrians. This is like asking how your automated car should behave when it is drunk; one of the great points of taking this out of human hands is that they aren't going to do the shitty things that humans do. Yes - you can jump in front of one going 60 under the right circumstances. You will die. This is not an ethical or moral issue for anyone, this is someone committing suicide. We've had automated trams and trains for a long time - how often does the old "do I divert the train to a side track to kill 5 people through action or let the train hit 15 people on it's continued course through inaction" problem come up for those algorithms? Never. It never comes up and is not part of them, because that is stupid. The train stops or the train hits the person the jumped right in front of it. I'm all for philosophical discussion, but you have to always bear in mind how little of this applies to real world problems.	Many think such cars will suffer from limitation as same as humans There is no accident without responsible Such cars will be more careful than human Such cars will not be drunk Could someone commit suicide by jumping in front of them We have trains and they should not swerve to save suicide people This is unrealistic situation	1 3	1.1 3.4
41.	Let's reverse the question. Suppose the car is on autopilot and there is only one occupant in it, the driver. Suppose the driver dies suddenly of heart attack or stroke: thus the car becomes "The Flying Dutchman." It continues to drive, carrying a corpse within. This may go on for hours, for days even, if the car has a self-recharge ability (i.e., it approaches known recharging points, plugs itself automatically, and has the payment charged automatically to the owner's credit account) and has been programmed to drive across the country, e.g., from Chicago to South Padre Island in Texas. A plausible situation, if the car is an RV and its sole occupant had retired to bed, prior to the stroke. How would insurance agencies, the law, police handle the situation? Should insurance (or credit card) companies hire special agents trained to identify and apprehend the Flying Dutchmen? If the car gets involved in an accident, who pays?	If a passenger died in self driving car and it continue drive from cost to cost, who is responsible? If car crash to something who is responsible?	3	3.5
42.	Maybe the most utilitarian approach, overall, is for automatic driving cars that sacrifice the owner if that minimizes loss of life, be mandated by government.	Utilitarian cars are good by law	1	1.5
43.	I'll never ride a car that is programmed to deliberately kill me - be it ANY circumstance whatsoever.	Will not buy self sacrifice car	3	3.1
44.	Sorry to say, however, this is a terrible article and is attempting to create fear amongst the uninformed public. The author is taking an antiquated question and applying it to a technology that he/she doesn't understand. These driverless cars run continual self-diagnostics to ensure it is operating within normal operating ranges and to ensure that there are no failures of the physical vehicle elements. This means the	This article create fear amongst public Situation is ununderstood Cars will not run if it has failure Designers will be careful Unrealistic	3 1	3.5 1.1

	engineers behind these products never have to contemplate the above issue because the car won't drive if it can't stop. I can understand the author's lack of understanding, because this technology is beyond amazing and the author may believe everything is just slapped together in a rush. This question was originally posed in the event of a physical mechanical failure of a tram in the early 1900's. Now, ask yourself, when a plane has undergone all of its engineering checks on a regular basis (remember this happens more frequently in these cars) do planes just fall out of the sky? Instead the question should be posed 'how do we engineer all products to ensure such a decision never has to be made'. This question is actually productive and can create results. But, I guess the author won't care, because their click bait has gotten us to read and engage with this article.			
45.	When cities are designed from the ground up for autonomous vehicles, or even if retrofitted for them, there will be a whole range of API's that will feed environmental sensor data, including where pedestrians are, the timing of the traffic lights, traffic congestion or other hazards, calculations of incoming traffic at intersections, etc to the vehicle OS. The on-board sensors will form part of a much larger data feed.	Such cars will be connected to the environment and run with predictability	1	1.1
46.	This is another article designed to create fear where fear shouldn't exist. Oddly, no writer's name is attached. But logically, by the time technology for autonomous vehicles goes mainstream, one could presume that the issues mentioned herein would have been addressed. Moreover, with the exceedingly poor driving habits people have, and the war-like attitude of most drivers on the road today, having a technology like this will obviously save lives. AND WE ALL KNOW WHAT HAPPENS WHEN BENEFICIAL TECHNOLOGIES THAT CAN EXTEND LIFE EMERGE- THE GOVERNMENT PUTS A STOP TO IT BECAUSE THERE'S MORE MONEY IN KILLING US THAN KEEPING US ALIVE, with exponential population growth. OK I'm finished.	This article create fear amongst public These problems will be solved Such cars will save lives	3 1 1	3.3 1.5 1.1
47.	I like the moral algorithm option ; The choice between saving ones own life and saving the lives of 10 at the cost his family's tears is a very personal one. Decision will vary from person to person situation to situation when humans are driving the car. There is no one right answer. I agree with another comment here that this is most probably an unrealistic scenario . However, should such a choice happen in a different situation, that choice should remain with the owner.	Unrealistic Owner could choose car rule	2	2.3
48.	This whole article is based on a false dichotomy: Kill others or kill yourself. The choice really is hit somebody or try to avoid them. You DON'T know if they are going to die if you hit them. You DON'T know if you're going to die if you hit a wall. With the advanced safety features today, there's a good chance you won't. This whole article is based on the assumption that somebody dies. In reality, we don't know for certain if anyone dies.	Unrealistic Is not self sacrificing or kill others it is hit or avoid Crash outcomes are unpredictable	1	1.2
49.	Early stages of Skynet. Distraction's inevitable. Just kidding. Why not stop the car or at least notify somehow the people that there's a	Stop the car	2	2.4

	problem. Assuming that the car makes calculations/observations per sec, how hard can it be to foresee an accident?!. The better the software the smarter the car, plus less possibility encountering an accident. (Variables and counters will pop off the charts though)	Warn and alert others		
50.	I see a flaw in this utilitarian approach. Lets say my friend and I want to kill a person under this system. All we have to do is jump in front of the vehicle when the victim is riding by himself.	Utilitarian is problem Will jump in front of car to kill its passengers	3	3.4
51.	Well you could solve the consumer dilemma for manufacturers by making it a user option. When a new driver is detected have a selection box visible on the display allowing the user to make the choice themselves, since when driving themselves that's what they'd have to do anyway, make a choice.	Owner could choose car rule Driving is same thing	2	2.3
52.	Amusing to examine this question, such fun! - Auto autos will result in vastly fewer road deaths, so they have enormous credit in that regard. - But, the universe being what it is, we'll surely get 'some' deaths, and 'some mistake deaths' and some 'hostile deaths'. It's just that they'll probably be few compared with now. - Auto autos will also be 'better than human-driven' simply because of their ability to decide faster (probably about 10,000 times faster). But, as with humans, they'll STILL make mistakes, and be 'hacked' mischievously. - We'll NEVER be able to get a 'perfect' set of algorithms for 'ethics' into the head of an auto auto, just as we can't get a perfect set into the head of a driver now. - In summary, hooray for auto autos. =)	Self driving cars will reduce deaths There will be victims but less than now Cars better than human There are no perfect ethics algorithms	1 1	1.5 1.1
53.	Let consumers make a choice, just like if they drive the car themselves. Some can chose to save the occupant life no matter what, and the others can pick the option of minimum loss. Also, people can change the option based on the situation they are in such as baby on board, pregnant wife, or passenger request.	Owner could choose car rule Option could be changed	2	2.3
54.	This scenario breaks down almost immediately: "...causes the car to head toward a crowd of 10 people crossing the road. It cannot stop in time but it can avoid killing 10 people by steering into a wall. However, this collision would kill you, the owner and occupant." A. The car can't know that hitting 10 people will automatically kill them (and its unlikely to kill all of them and may only cause injury.) B. The car can't know that hitting the wall will kill the driver. Nobody knows these things -even ordinary drivers won't know the outcome of a collision, one just takes one's best shot at minimization of injury or death. A self driving car should simply come to a full stop in any situation where it doesn't know what to do -same with all of us. The real issue is: Who is liable? The driver? She/He didn't program the car. The software developer? That person isn't at the scene. Answer to the whole problem? Self-driving cars are as bad an idea as self-shooting guns. They should never be allowed. Lots of things can be made better with automation. Driving is not one of them.	Crash outcomes are unpredictable either hitting pedestrians or hit wall Even driver can not predict the outcomes Self driving car should stop The problem is responsibility Self driving cars like autonomous weapons shouldn't be allowed	1 3	1.2 3.2

55.	I really don't see why such things are thought to be even as straightforward as are commonly put forward, anyway. What if it is 5 old people who are clearly with one foot in the grave, and a baby in a baby carriage and their parent. What then? What if it is those 5 geriatrics and then triplets on the sidewalk with their parent? If I had time to think, I would choose the geriatrics in both cases for demise, but there is no way any time soon that the self-driving cars can make any such kind of assessment. And I guess I am saying that the short number of combined years for the geriatrics is "worth less" than the many total years represented by the baby(ies) and the parent. It is worth noting that many people would accept my assessment and many would not. Either way, the idea of self-driving cars will be blamed for moral failure and as such will simply never be allowed. It is too easy to scapegoat the technology as a negative thing, even if in general it could drastically cut traffic deaths over all.	Will choose old people No way that self driving car will calculate people ages This is moral problem will slow down the self driving cars idea Even they will reduce total deaths	3 3	3.2 3.5
56.	this is utter bullshit. What would you have done if you are driving a car heading towards a 10 people crowd with high speed? In this extreme situation there will be loss of life without doubt, even it's not a self-driving car. Real life drivers will give the same exact reaction of steering away from the crowd and hit the wall, so are human drivers programmed to kill to?	Unrealistic Human drivers will swerve to avoid pedestrians	1	1.2
57.	How was there a group of people a short distant a way that the car did not recognize? Was the car driving too fast, or did the group of people jump randomly in front of the road. If the car was going to fast, that is an engineering problem that could be avoided preemptively. If the car was turning the corner, it should not have gone that fast around the corner, there would need to be programs to avoid that. If it was foggy, the car needs to drive slower in the fog. If, by chance, a group of 10 people jumped in front of a speeding car, then we have a game of chicken, and we can not say the car is at fault. The people acted maliciously. Though, I would like to take this thinking to what I believe is a more controversial topic, autonomous weapons. There are laws of war that protect civilians, and I am wondering if it is the technology that kills civilians if the engineers should be responsible for the war crimes. Is it possible to create technology that does not kill civilians.	If the car drives fast it is programmers problem The car should be drive slowly to avoid crashes If people jump in front of the car it is not the car fault It is possible to create machines don't kill civilians	3 3 2	3.5 3.4 2.2
58.	How would the car know that what it is about to hit are people as opposed to animals, objects thrown on the road, etc? Also, how would it know that what it will veer into is a wall that is hard vs. a guard rail? I'm curious to know if this is part of the foolproof algorithm before the moral dilemma becomes an issue. Otherwise the problem would just default to protecting the driver no matter what objects are in front of it	How the car know other objects? Maybe animals or thrown stuff Are these computed? Protect passengers	3 1	3.4 1.3
59.	This tired, contrived hypothetical again? If you are going to give the car near-omniscient ability to tell whether or not its actions are going to result in a fatality or rank the injuries that will result from its actions - and make value judgments based upon that - then you might as well give it the credit of being smart enough to not drive so fast that it can't avoid obstacles that are	If you want to give such cars ability to decide, praise them being good Such cars will behave like human driver	1 3	1.1 3.4

	beyond its ability to detect. Whether you mean the car itself or the people who program its behavior. You know, like how human drivers are not supposed to drive so fast and so close to to the cars in front of them that they can't stop in time if the car ahead hits the brakes. I'm going to need to see an example of one of these "unfortunate set of events causes the car to head toward a crowd of 10 people crossing the road" before I believe it will happen in real life despite the cautious behavior they will need to be programmed with in order to get declared road-worthy. The whole point of autonomous cars is that they won't be doing the stupid things that humans do, because they're bored or impatient, that cause the accidents. As drivers, humans really are not that good. The real "unavoidable accidents" are going to be ones where something like the road surface suddenly fails underneath the car, or a tree or boulder falls on or immediately in front of the car. Rather than dealing with overthought ethical judgments, the cars will do the safest thing they can: come as close to a stop as they can and minimize the energy involved in any collision so there's the least amount of overall damage. There's no moral dilemma involved in driving a car, only physics and humans in denial of it. This is just underemployed "ethicists" trying to convince people they are relevant to an issue they aren't relevant to.	Self driving cars are better than human drivers The real problem is something thrown in front of the car The self driving car will stop There is no moral dilemma Underemployed ethicists try to be involved Unrealistic		
60.	The car should protect its owners at all costs, why are pedestrians in the street in the first place? If they erred and failed to move the people in the car should not suffer. Easy solution.	Save passengers at all costs	1	1.4
61.	no way I will buy a self driving car, there is always an alternate decision a human can make that a computer isn't programmed to make	Will not buy self driving car	3	3.1
62.	why doesnt the car just stop	Just stop	2	2.2
63.	Exactly. Hello? Just stop. Throw out the anchor, lock up the binders, and s-t-o-p. Default position. End of ethical dilemma.	Just stop	2	2.2
64.	Very interesting..out of the top of my head I'd say, let the car owner decide what kind of algorithm is preinstalled in their vehicle, the "sacrificing the passengers type" or the "killing the unfortunate pedestrians type" and assume legal/ethical responsibility accordingly. It all comes down to "are we trully utilitarians when our life is at stake?" Are we the perfect Socratic type of citizen (who btw, was quite old when he took the decision to end his life at the command of Athenian democracy) or we simply pretend to be until our own head is at stake?	Owner could choose car rule to be responsible	2	2.3
65.	This isn't even a real problem. The car/computer/set of sensors is/are better equipped to avoid these situations then people are. By a lot. I had this come up in a philosophy class, really this is just a glorified trolley problem/dilemma that is nothing more than a thought experiment. It is as well rooted into reality as the ethical dilemma of whether or not it is ethical to kill an alien which has visited earth if it plans to experiment on you. (forgive that horrible example, I am intentionally being outlandish, like this experiment).	Unrealistic Such cars better than human It is just thought experiment	1	1.1
66.	This scenario would suggest that the hazard awareness system isn't up to the job. It's also highly unlikely of you think about how hard it	Unrealistic	1	1.2

	would be for one car to kill 10 pedestrians unless it was speeding through a crowded high street. A much bigger and more feasible problem is the autonomous nature of these systems. It's all very well one system deciding on the best course of action in a given situation, but it's no good if two cars decide to pull into the same space, for instance if cars in lane 1 and 3 move into lane 2. Systems will need to be aware of each other, and make group decisions. A more interesting situation is a sudden accident ahead where cars need to avoid a pile-up. With a hive-mind system cars could work together to prevent collisions, as well as warning cars behind to slow down.	What if two self driving cars change their lanes to one lane?		
67.	Clearly this is ridiculous. What if it's 5 deer and the driver dies for deer? What if they car swerves so fast that it flips and is an even wider death machine while killing the passengers as well? What we need is equipment that stops all texting and driving	What if animals jump in front of the car	3	3.3
68.	This is what road side units and network infrastructure are for, these will detect crossing pedestrians earlier than the car can, communicate that to the car which will slow down before such a scenario occurs. That's the idea anyway - it is not only the car involved.	Such cars will communicate with environment	1	1.1
69.	I don't fully understand. As long as the vehicle did not do a mistake, why should it's occupants be blamed or injured? If someone (or even a group) crosses the street where he/they should not, then that as consequences. That's the same as with normal driven cars. You also cannot blame the driver of the car. Of course it would be better to only injure the driver instead of killing someone who made a little mistake, but that's an optional add-on. If someone doesn't obey the rules, why should he not pay for it?Or, another very different point, what if a pedestrian intentionally jumps on the street directly in front of an autonomous car? Will this be the new type of mother-in-law murder?	Why blame passengers if someone jump in front of the car It is optional to injure passenger instead of killing someone Jaywalkers must pay their mistake	3	3.3
70.	How about we talk about assassination. I'm actually surprised this article doesn't even mention it. How strange is that huh? Hmmm. Well, we all know that EVERY type of programmed device can be hacked.. so, what about individuals or organizations who decide to kill people intentionally? This is really no different than a 'hacker' changing the programming in a robot to go rogue and start killing people or it's operator. This type of scenario is VERY possible, no real debate needed. The history of technology has proven itself to us all. Most of us know what the darker side is capable of, unless you take us for complete buffoons. To assume any different only underestimates the intelligence of the average person. Do you really think most people are that stupid? Think again. The control grid isn't going to fly, at least not for very long anyways. There will be a huge opposition, especially as intelligence grows pertaining to how the inner workings of technology.	What about hacking such cars to kill others	3	3.5
71.	I think there's no moral dilemma actually. Self-driving cars only should take care of its occupants. Just like the ordinary cars. The driver, instinctively will take care of himself / herself and the passengers over other people who eventually could be killed. And nobody hesitates about the morality of this behavior. Besides, a driver could kill all the occupants in the vehicle trying to avoid running over a pedestrian (for example) , so there's no reason for considering that human reaction will be better than automatic decision based on save primarily the lives of the occupants. And	Save passengers No one will be self sacrifice cars Self driving cars will reduce accidents	1 3 1	1.4 3.1 1.5

	finally, if a potential buyer of a Self-driving car knows that its future vehicle could decide killing him or her... who wants it!? No market for this product. There's accidents. There will be accidents. But if all the cars are Self-driving cars I'm completely sure that those accidents will be clearly lower than today.			
72.	The real question is how two self-driving cars would fare in a game of single-lane, head-on chicken, coming out of a blind corner on a tight road!	What about two self driving cars in same situation	3	3.5
73.	This makes killing drivers rather easy. Just take your friend with you, search for a narrow point, wait for the car and both of you step on the road in front of the car. You know, the software has to kill the driver. Stupid thing. Just an idea: Maybe "minimizing the death toll" is not the best solution. If - for example - a car has to choose between hitting a person on the road or hitting a wall then it should protect people in their "safe zone" (which is where they are if everything is normal; i.e. a road for a car or a pedestrian on the sidewalk) which - in this example - would be the car because the pedestrian on the road already left his "safe zone". Another example: If one pedestrian is walking on the sidewalk and 10 people are on the road, the ethical solution is to sacrity the 10 people on the road because they already left their "safe zone".	Unrealistic Jumping in front of the car will kill the passengers Minimizing deaths is not a solution Do not sacrifice innocent people	1 3 3 3	1.2 3.4 3.1 3.3
74.	What scares me the most is that no one here has asked the most obvious question of all: Why are there 10 people jaywalking in the first place?	Unrealistic	1	1.2
75.	The car should be able to realize that it is about to head into a wall to sacrifice itself, and at that split second eject the driver out of the vehicle (seat belt and door opens at the same time) or a massive air bag can open which fills up the entire space in front of the driver just as the car is making its self-destructive decision.	Eject the passengers and drive to wall	2	2.2
76.	I think that if the dilemma happens, if you kill one or 10, you go to jail. If by avoiding killing them you crash yourself against a wall, you're dead. There is not much difference. Now, on the other hand, I figure if I'm driving a 'normal' car and I have 10 peoples crosing the road, I woud try to avoid them, but my mind can not reason in a split of second if the fact of avoiding them will kill me, and therefore, being performing an act of self-sacrifice on behalf of 10 people. I imagine myself trying to avoid and that is what God wants. Why I will benefit from the speed of calculation and response of a computer that possibly, to avoid me to crash into a wall and die, it ends up taking the decision I had not taken being in charge?	Killing people or self sacrifice is ethical dilemma	3	3.5
77.	I think the car should run over the 10 people. If the car is automated to abide by every street rule, then whenever and wherever the car is passing, it is abiding by the rules. It means it's also driving at the right speed limit to have time to break properly if an obstacle appears. So if 10 people jay walk in a split second, they're in the wrong. They should no better. It's actually unfair to the drivers. You're not supposed to jaywalk. You're supposed to look right and left a million times before you decide to jaywalk. Imagine a world where jaywalkers don't need to do that anymore	Jaywalkers must suffer Such cars are abiding the law If car avoid pedestrian, jaywalkers will jump if front of it This is not fair	3 3	3.3 3.4

	because they know that self driving cars will avoid them to kill the driver. What if stupid teenagers start jaywalking on purpose to mess with self driving cars? What if criminals starts taking advantage of this as well? What if you're in a dangerous situation and you can't get out of it because someone is simply standing in front of your car? I think we need to respect the car. If all cars are automated to be in the right and you still manage to get hit somehow, then you must have been doing something wrong. Obviously there are always exceptions, maybe someone fell to the streets by accident, but those are freak accidents and unfortunate.	What about bullying and taking advantages Jaywalking maybe wrong (illegal)		
78.	The hypothetical postulated here is: A self-driving car suffers a complete brake failure while approaching an intersection at speed. This is a situation that is incredibly uncommon with conventional automobiles, and is likely to be even rarer with self-driving cars, which remove many of the opportunities for bad judgment and neglect that would cause this problem to occur in the first place: First of all, owners of conventional automobiles can neglect proper maintenance of their cars, which might result in brake failure. A self-driving car can take itself in for service and it can be programmed to refuse to operate if its owner has delayed important maintenance to the point where the car poses a risk to riders or pedestrians. A car driven by sophisticated computers may also be able to run complicated diagnostics on itself that current cars cannot, or detect problems before there is a complete brake failure. Second, a self-driving car will drive the speed limit and brake at a safe distance from an intersection. If it experiences brake failure, it may be able to honk its horn and trigger its emergency brake to give the pedestrians time to clear the intersection. If the people run into the intersection without the right-of-way, and the car can't avoid them without imperiling its owner, then it should hit the pedestrians. The party that is acting wrongly should bear the risk of harm that accrues from their actions, and if the cars are programmed to swerve out of the way of people who run in front of them, then criminals may exploit this rule by running in front of cars to force them off the road. But the most common situation in which this might occur -- children running into a street -- is still less likely to lead to tragic consequences with self-driving cars, because residential streets and school zones have reduced speed limits precisely because of this possibility. Human drivers often disregard the speed limits, and autonomous cars will not.	The real problem is break failure Self driving cars will refuse to run if there is failure Such cars will obey law and if there is a failure it will alarm and alert If pedestrian jump in front of the car, it will hit them Jaywalkers must suffer If the car programmed to avoid killing, criminals will hump in front of it	1 2 3	1.1 2.4 3.4
79.	Just make it optional and allow the owner to set his desired operation.	Owner could choose car rule	2	2.3
80.	Wouldn't it be perfect if the car could also detect what is exactly in front of them from a certain distance, for example it could tell if it was a car or a group of 10 people or even a bird or a cow, this way based on the distance, the car could have enough time to slow down based on the speed the objects in front of it is moving. So technically, the solution would just simply require an increase to the	Cars will detect pedestrian and stop	1	1.1

	cars vision and programming speed? I guess it would take a lot to develop something like this but who knows, automated cars are already here.			
81.	The example has no moral dilemma. The driver of the car should be sacrificed because she chose to use a deadly machine. Ordinary people cannot wait to see the day that drivers start to pay for the deaths and misery they cause. The real scandal is the immorality of the present system, where drivers impose a heavy cost to everyone else.	Passengers should be sacrificed because they decide to use the car	3	3.5
82.	an automated car should be built in order to do better what the owner would, so the real case is that the driver in such situation will look to diminish damage as long as his life is not affected, and that should do the robotic pilot also.	Cars will do better than drivers to reduce damages and save passengers	1 1	1.1 1.4
83.	The car should be fair, it should respect the signs and the rules, if pedestrians (doesn't matter how many) are not crossing the road at the right place, then why the car should kill the innocent owner?	Jaywalkers must suffer Shouldn't kill innocent driver	3	3.3
84.	We expend much engineering effort improving the crash performance of vehicles for the benefit of the occupant, but relatively little engineering effort decreasing vehicles' aggressivity indecies to protect pedestrians. I have never heard of airbags on the outside, but would suggest this as a first option before we decide to program vehicles to kill vehicle occupants via suicidal evasive manoeuvres. Sensible land use and transport policies also go rather a long way to reduce road deaths, driverless or not.	Such cars should protect pedestrians by deploying outside airbags	2	2.2
85.	Being someone who designs autonomous systems, I can say that these systems are presently not so sophisticated that they can even detect the difference between any obstruction (another car) and a pedestrian. Secondly, this article completely glosses over the fact of how the car got into it's ethical dilemma. Autonomous cars can 100% be programmed to control their speed in such a way as to make this kind of situation incredibly rare. Say a car can only see 100 ft in front of it (due to a hill, darkness or other factors). Say the same car can stop from 60mph in 100 ft. In this case the car should never go faster than 60mph. Problem solved. Also, a 5000 pound chunk of metal (a car) can not magically change it's direction on a dime even with a computer as the driver. Who says this problem has to be solved before cars can drive themselves? Current generation cars can not even detect this situation exists, not to mention making a decision about it. Nevertheless they control the steering wheel. Also, how about we only allow (at first) use of autonomous systems on the highway (where it is illegal for pedestrians to be) until these factors become better understood. Why do we have to build a system with the complexity of a rocketship when all we need is one with the complexity of well... a car.	Such cars will be programmed to follow speed rule They will not change their lane at all What about using them in highways No pedestrians there Start with that until we solve other problems	1 2	1.1 2.1
86.	The trolley problem needs to die. These kinds of puzzles are artificial constructions designed to highlight a tricky moral dilemma but have little applicability to the real world. The situations are so unrealistic and so unlikely to happen that they don't reflect real life at all. Why are there 10 people in the middle of the street? Why aren't they paying attention? Why didn't the car's sensors see them	Unrealistic There is no way to have life and death situation Crash outcomes are unpredictable	1 1	1.2 1.3

	sooner? The scenarios are also artificially framed in tragic, binary, life-and-death terms. If a machine chooses to crash into a barrier to avoid hitting a cyclist, it's not necessarily choosing to "kill" it's occupant. Many people survive even high-speed crashes, due to seat belts, airbags, crumple zones and collapsible steering wheels.. It's like saying that a manager who must choose who to let go and who to retain is choosing who to condemn to a life a poverty. That's certainly one outcome, but the newly unemployed person may also end up finding another job, or even a better job.	Cars have safety features		
87.	I am not a lawyer but if I was one I would be collecting everything I could that could later be used in court to argue about deciding about the programming of who has to die. I could argue that the pedestrian shouldn't have been chosen if the pedestrian died. I could argue the next case differently. In the Jag TV shows the lawyers would be handed cases and sometimes they would be the prosecutors and other times the defenders. It worked. Driverless cars will make lawyers money because families of everyone who dies could bring a lawsuit. The programming would always be the placement of the blame. I am also considering investment options in the companies who hold the patents on things related to the technology because driverless cars are going to reduce car accidents to almost none.	Lawsuits will increase because of self driving cars problems Programmers should be blamed Such cars will reduce accidents to non	3 1	3.5 1.5
88.	I presume no one has thought of a complete shutdown of the vehicle save for a signal to surrounding cars of such a maneuver? And how may of the F1 and Indy 500 drivers have been consulted for strategic maneuvers that could help out as well? Example: while driving on a closed street course during a practice once driver found his brakes failed while approaching 90 degree turn, threw his car into a sideways slide and made the turn. In this case the car didn't hit the building, not so sure that would be the case with all professional drivers but it shows it can be done. Quite sure the computers would react faster...	Self driving cars will react faster than human drivers	1	1.1
89.	I am curious who will authorize the software used in these self-driving vehicles. Late emission test frauds in one of the biggest automotive company implies that "everything goes"-attitude is acceptable and shortcuts will be made.	Who will authorize self driving car	3	3.5
90.	There is at least one additional moral wrinkle to consider. Hitting the pedestrians is nearly certain death for them. Hitting the wall is not certain death for the vehicle occupants unless the car is traveling at a very high speed, as they are protected by safety features of the car (crumple zones, side impact rails, air bags).	Safety features in cars will save passengers so hitting a wall is better than hitting pedestrians	1	1.3
91.	So, if the car confuses sheeps with humans, will it crash the car and kill the occupant(s) to save the sheeps? No way, if you see an automated car coming ... run!	Will such cars confuse between animals and humans? If yes, run	3	3.1
92.	I would not ever risk the safety of myself, my passengers, or an innocent by stander on the sidewalk because someone crossed the street and was not paying attention. Picture A could easily represent the group of protestors in Oakland, CA who blocked off the road. I would not hit an innocent person on the sidewalk over them.	Will not sacrifice self or passengers or innocent people because of jaywalkers	3	3.1
93.	Just a quick thought. Instead of thinking the "dilemma" way, why don't we think the "problem solving" way ? That is to say, to avoid any of this to happen, why don't we include ejector seats in all those car models.	Eject passengers outside the car	2	2.2

	Thus, the car could always behave in a way to save both people outside the vehicle and the owner (+ other passengers) himself (themselves). And we can not say this would be too costly (anyone able to buy this kind of car could add a few dimes to add the life saver seat)			
94.	It's clear we need pedestrianless pedestrians. Automated pedestrians. Then no one walks or runs out in the street without first knowing if that driverless car is going by.	No one should walk if there is self driving car	2	2.1
95.	"And therein lies the paradox. People are in favor of cars that sacrifice the occupant to save other lives—as long they don't have to drive one themselves." Weird how this could've been said without further thought, this is a known aspect in psychology. It's like soccer; when your favorite team wins you would say: 'we' won! if they would've lost you would have said: 'they' lost. This means we do not associate ourselves with the losing party, we would physicly and mentaly move out, which means it is a whole different starting point. The other thing I can think of (as a System Engineer myself) why would we even want to solve it? why do we even want computer controlled vehicles? Maybe we could better put our time and money into better alternatives? improve or replace public transport, run computer controlled and human controlled (vehicles) apart from eachother (like seperated ways) and maybe look some more science fiction, they already thought through some of these issues.	It is natural for participants to say sacrifice passengers but they don't want to be in the same situation Why we lose time to solve such problems We could enhance current cars	3	3.5
96.	This article illustrates why no one subscribing to utilitarian ethics should be allowed to program self-driving cars.	That is why we do not want utilitarian cars	3	3.1
97.	The purpose of ethics is to identify general principles of action to guide one's actions over the course of a lifetime for the purpose of living a happy and productive life. One cannot "test" an ethical system against contrived thought experiments where one must make a split-second decision among bad outcomes. This is a favorite mind-destroying exercise of philosophy professors and other cynical cranks.	Unrealistic Why we have such thought experiments	3	3.5
98.	If a car that drives itself can't stop in time to avoid killing anyone then we got a serious issue, it's 2015, I've personally experienced a BMW going 60-0 in under a second, and it had a human driver, now I Imagine a car that drives itself is smart enough to actually stop when it senses danger. Let's put it in a real world situation, you're cruising trough a major city, the streets are going to be packed pretty much all the time, streets where you would meet crowds of people would be closed most of the times anyway, there are stoplights, there are speed limits, and I'm sure that an automatic vehicle will respect those, I understand the concept of abstraction, but IMO this thing is senseless. Wouldn't it be better for one to give that car some sort of manual override in the extreme case situations like this arrive? Wouldn't self-driving cars be a bit dangerous to start with? We don't need a solution if we don't create a problem. What if we just stuck to CC and automatic gearboxes, I really think you shouldn't own a car if you're too lazy to drive it.	Manual driving is good option in such cars Shouldn't own car if you do not want to drive it	2 3	2.3 3.5

99.	What if it is a purposeful hijack situation where a person or persons with weapons, intent on taking the car or occupants of the car by force..or killing them? How will the car handle that?	How about hijackers	3	3.5
100.	A better question is do car manufacturers and insurance companies have the balls to let driverless cars become ubiquitous. By building a driverless car , the car manufacturer pretty much takes over ALL the responsibilities of harm that the car can do. The owner is now just an innocent passenger, so if the car ever gets into an accident, the manufacturer will have to pay for damages of BOTH parties. Even if driverless cars can reduce the amounts of accidents by 90%, that's still a lot of potential payouts. The $30k price tag of the car will likely not cover the amount of "accident" money that the manufacturer will have to pay over the car's 15-20 year lifespan. As for auto insurance companies, they will go out of business because owners will no longer need to buy insurance as it is the manufacturer's responsibility.	Manufacturers should take all responsibility Passengers are innocent In accidents, manufacturers should pay for both parties No need for insurance since responsibility on manufacturers	3	3.5
101.	You have a few problems here. I understand this is a hypothetical situation, but why is there a crowd of people in the first place where there is no controlled access (a smart car would be aware of where crossings were so would not likely be traveling so fast as to threaten them) and why should a driver be harmed while pedestrians are free from the consequences of being where they don't belong or crossing against a signal in a controlled intersection? If you design smart cars with pedestrian avoidance, human nature would be to expect the car will simply or swerve and crash leaving them to cross where ever they felt. If smart cars are going to work as a concept, they either need to only be manually driven in the city or stronger enforcement/stiffer penalties against jaywalking need to be created.	Unrealistic Why punish passengers because of jaywalkers Jaywalkers must pay penalties	1 3	1.2 3.3
102.	In my opinion the car should try to minimize the damage for both the occupants of the vehicle and the people on the street. It should slow down to a stop as soon as possible without swerving in to a wall or down a cliff. Makes no sense that the car should sacrifice the occupants when it is the crowd walking over the street that most certainly breaks the law. What kind of street is it by the way? This hypothetical situation seems very unlikely to happen. A crowd of people crossing the freeway, makes no sense? If it is in the middle of a city the speeds should be so low that the car would most likely be able to stop or slow down to a speed where deaths could be avoided.	Reduce damages in general No swerving, just stop Why punish passengers because of jaywalkers Unrealistic	1 2 3	1.5 2.2 3.3
103.	Pretty simple really, 'Thou shalt not kill' give up playing with technology that will make that kind of decision. Solution – hand the responsibility back to the human.	Human is responsible Stop playing with technology	3	3.5
104.	So what if some of those little twerps who drop big rocks off of highway overpasses decide to get together with a few of their friends and run out in the road, "for the lulz?" What are those ten people doing in the road anyway? Give the driver control.	What if someone jump in front of the car Let driver decide	3 2	3.4 2.3
105.	False dichotomies. The spectrum of choices in real life will be far more fuzzy and ambiguous than the clear, almost-either-or choices constructed artificially here. Almost certainly, the vehicle's sensors will allow braking (or, if the brakes have failed, other evasive maneuvers) to remove kinetic energy long before the point of no return imagined here, and allow for high probabilities of injury instead of death. Also, given airbag tech advances, almost impossible to imagine a scenario where the occupant doesn't survive. BTW, lawyers have already been working on these issues for nearly a decade. And the realm of their inquiry involves not only	Unrealistic More choices than two The car will stop If failure, will slow to injure not kill Cars are safe for passengers	1 2 1	1.2 2.2 1.1

	autonomous vehicles, but autonomous warriors (law of war implications).			
106.	The answer is to protect the occupant of course. This is a no brainer. Jaywalkers are type of human filths that not only put themselves in danger, but the driver and other people around them as well. Any worthless jaywalkers who would even cause such life or death dilemma should consider their lives forfeit.	Protect passengers Jaywalkers must suffer	1 3	1.4 3.3
107.	"And therein lies the paradox. People are in favor of cars that sacrifice the occupant to save other lives—as long they don't have to drive one themselves." It's not a paradox. People simply framed the question as "provided self-driving cars are novelty thing I have nothing to do with, it should do X". Of course everyone's going to make a generalized statement that minimize the death toll. The question should be set up as, "You drive autonomous car X which is about to either kill you or kill 10 people. Your son is in one of the 10 people." and so on, so that it is forced that the people would have skin in the game. Then watch what opinions unfold.	It is natural for participants to say sacrifice passengers but they don't want to be in the same situation Pedestrians could be relative of driver How to act?	1	1.5
108.	As many people here note, these are highly constructed situations presented here. We probably should not be bothered by this fact, it is about one important distinctions between humans and machines: we are very willing allow humans to interact with the world without having to state their rulesets. We require people to follow general traffic rules, but this will be incorporated in any autonomous car, which will follow these rules way more exactly. We are talking here about rulesets "rather evade than brake" or "avoid hitting the hard targets", which human drivers need not declare before being allowed to drive and they probably couldn't, as they will most probably not even know them themselves. Machines, on the other hand, are quite fixed on these rulesets, as we need to program them into the system. So, because we have these rules on paper we start to think about them. We could simply ignore this knowledge, which would be more fair, as humans don't have to give us their rulesets, but this is probably not ethical. With machines, we have the rules and we need to decide what would be the best rules. General traffic rules are easy, any reaction rules for uncritical situations are also easy. As machines are so much better than humans many situations which would be critical for a human will be solvable by an autonomous car, which leaves us with very critical situations, those which look constructed, those which are so fuzzy in reality that they have to be simplified to be useful for any ethical evaluation. We are able to define rules even for these strange situations, and for this experimental ethics might be a good approach. We, as humans, sometimes think about wrong decisions after the damage is done. Here we enable us to decide what a good decision is. Call me utilitarian, I think this is good.	Such cars will follow the law Declaring rulesets is unneeded from human drivers before license and they couldn't figure it We have rules, we decide what best and teach machines Machines better than us	1 2	1.1 2.2
109.	Scenarios described in the article relate more to human drivers where a lapse in attention has caused a situation requiring a split-second decision. Realistic scenarios like that involving a self-driving car will be rare simply because they will always obey the rules of the road (e.g., speed limits), they don't get distracted, and their reflexes will likely be far superior to a human driver. Accidents involving self-driving cars are more likely to result from the system failing to properly recognize what it "sees".	Unrealistic This scenario for human drivers not cars Such cars obey rules Accidents because of failures	1 1	1.1 1.2

110.	As many have pointed out, the scenario presented is unlikely: given an autonomous vehicle travelling on a city street, the speed limit is set that a vehicle should have ample time to stop when encountering a single pedestrian or more in the roadway. The other part that makes this scenario unlikely is the assumption of death: as the probability of encountering a group of pedestrians in the roadway increases, the probability also increases that the speed limit will be low enough that steering the vehicle into a wall or other rigid feature will not result in driver casualty (i.e. the speed limit in most cities is 15-35mph, which should not be fatal to vehicle occupants). Someone is thinking way too hard about this while sitting at a desk. If you want to know how people will respond so you can program accordingly, don't survey them (which is subject to bias). Base it on how people have ACTUALLY RESPONDED when presented with the situation by using actual crash data. In the real world when the situation is encountered, the driver has a split second to react, and may do so instinctively. There is no time to think through it, to wonder if the pedestrian has children to go home to or not, to weigh the value of your life, et cetera - there is only time for reaction, and sometimes, not even that. So in the real world, what have real people done? Again, it is an unrealistic, worse case scenario that is easily avoided every day by millions of alert drivers driving defensively.	Unrealistic Crash outcomes are unpredictable Bias in choices come from such surveys	1	1.2
111.	I would never have one. What if the computer fails (and it will)? People will be come lazy behind the wheel. Humans should NEVER be replaced completely by computer when it comes to things that can take or save human life.	Will not buy one Machines fail Humans shouldn't be replaced by machines	3 3	3.1 3.2
112.	Another strong indicator that we should invest in public transportation instead.	Doesn't need such cars	3	3.1
113.	There is a major oversight in this dilemma. A self sacrificing car that reacts on certain variables beyond the control of the driver, could be tricked into committing suicide. Pranksters could attempt to cause an automated response to crash a car. Even choose a location where the options were impossible for the car to avoid destruction. A dead hard stop and limited swerve is the best AI can call for. Self sacrificing is not an option at all. That must be an act limited to the driver.	Self sacrificing is suicide Pranksters could take advantages Driver must act	3 2	3.3 2.3
114.	Very simple, eject the passengers and use drones or wathever device (hydrogen powered jet/explosion inside seats) to keep them hovering and provide safe landing while giving enough time for other drivers to safely avoid. This only needs a couple of seconds, so power needs aren't huge. If we can build drivable cars, then find a way to eject the pilot safely, it's not that much harder. After this, do wathever you want with the car: crash it, vaporize it, split it, flip it... I'm sure there are other ways to avoid suicide or kill decisions, so I don't think we're going in the right way.	Eject passengers Don't kill	2	2.2
115.	The issue gets more complicated. Suppose the group of 5 people are reckless jaywalkers and the one person on the side is a careful pedestrian staying on the sidewalk; should the one person be sacrificed?	What if they are jaywalker? Sacrifice innocent one?	3	3.3
116.	If You encounter such situation while you DRIVE the car what do you do? I think the most people protect themselves and kill the others though it is ethically wrong.	Human drivers will protect themselves humanitarianism will reduce deaths	1	1.5

	Maybe a humanitarianist minimizes loss of life. That is, It is just dilemma. it cannot be solved as ideal. Currently, when such situation comes to human driver, the effect is random. It depends fully on driver's behavior. So the best alternative solution is to make the case better than when human drives according the statistics of such accident or whatever.			
117.	See the problem with ethical problems like this is that they fail to look at the whole picture. They assume absolute outcomes where the AI will KNOW with 100% certainty whether or not an action will result in death. Not even people know these things with such certainty. It's about balancing probabilities not certain outcomes. Modern cars are packed with air bags and designed to crumple thereby reducing the damage to occupants in the event of a collision. Modern pedestrians are not equipped with such features. Thus: All other things being equal, drivers and passengers are much more likely to survive a collision than pedestrians. Vehicle AI should thus have a policy of prioritizing pedestrians even over their own drivers and passengers. Vehicle AI should prioritize their own drivers and passengers over those in other vehicles, however. Only exception to this should be where children are involved... assuming vehicle AI can't tell if another vehicle contains children, then at the very least it should prioritize school buses (particularly during school hours or in school zones) over their own driver.	Crash outcomes are unpredictable not even for human drivers It is probability Cars have safety features will protect passengers while pedestrians have non Such cars should prioritize pedestrians And prioritize passengers over other passengers Prioritize kids, if not possible prioritize school bus	1 1 1	1.2 1.3 1.5
118.	The ejection seat with a parachute solves this problem. Eject the occupants of the car, car crashes into wall. No one died.	Eject passengers	2	2.2
119.	These cases are very rare. It doesn't happen all the time. Even if so, there can be a solution for this. If the AI is 100% sure that there will be death, it can deploy a mechanism to stop it from happening. This mechanism can be something like to drop an anchor at the rear of the vehicle into the road, it can bring a car into a dead stop. Although the driver can be injured during this process, the damage can be reduced by immediately tightening the seat belt, deploying air bags (and/or any other method to prevent the head to hit the steering wheel or the window) and dropping the anchor, all at the same time. Or dropping the anchor 5-10 milliseconds after deploying the driver protective measures. Before any of this happens, all the interiors of the car can glow red as a warning to even give 1-2 seconds for the passengers to prepare themselves. If you're a driver, then you should know 1 second or even less than that is A LOT ! The anchor could be anything, be it purely mechanical type as in dropping a huge nail sort into the road which will be attached to the whole car, or it could be a strong electro magnet. And of course, the process to do all this has to be super fast and it should be deployed only when there will definitely be a death involved. So the AI has to be super fast and super intelligent. If an AI can be created for that, then I think it would be a good idea. Also, fixing the road will be much more cheaper and feasible than bringing someone back to life.	Unrealistic (rare) Such cars are capable to act safely Safety features will reduce damages to passengers Warning for passengers to prepare Dropping anchor is good idea	1 1 2 2	1.1 1.3 2.4 2.2
120.	If the car is heading to a crowd crossing the road in a situation that it will potentially cause an accident, it means that these people are crossing the road at a time and/or place where they shouldn't, because otherwise the car wouldn't be going fast enough to be dangerous. Either that or the car is malfunctioning, and in both cases it can't be programmed to kill me by someone else's fault.	Pedestrians brake the law Such cars will not kill passengers because of pedestrians fault	3 1	3.3 1.1

121.	"One day, while you are driving along, an unfortunate set of events causes the car to head toward a crowd of 10 people crossing the road." An unfortunate set of events? Is this supposed to be science? If you obey traffic laws, then that should NEVER happen. There is no circumstance I can possibly imagine where your car is suddenly headed towards a crowd of pedestrians faster than you can stop it, except jaywalking. If they are jaywalking, the driver certainly doesn't deserve to die for it.	Unrealistic Except jaywalkers In that case, passengers shouldn't die because of them	1 3	1.2 3.3
122.	Two issues that seem to be missed in this old debate: 1) outcomes involving human reactions/behaviour are rarely predictable and controllable 2) rules influence future behaviour 3) bad rules can be gamed (all 3 options open the door to a perfect crime) With option A,B or C, you're removing the autonomy of the road-crossers. It might seem cold to use put a principal (autonomy) in front of human lives at stake, in one scenario. However, if we're discussing principals and not just values, we're responsible in this decision not for one, but for thousands of such scenarios repeating over time. Ask yourself honestly: If you knew that you were no longer responsible for yourself when crossing a road, would you - consciously or not - cross roads with quite the same sense of caution? Or if you would, can you say the same for everyone you know? Option D) treat humans as mostly-autonomous beings; break and swerve but not to the point of putting the passengers at risk, as a typical human driver would do. Would this lead to the best outcome in this scenario? I'm not sure. And I don't trust anyone who is. Would this lead to the best outcome for thousands of scenarios over time? So far I think so; by following human nature, and reducing unwise road crossings to begin with.	Crashes outcomes are unpredictable Options could be used for crimes	1 2	1.2 2.4
123.	What is a cop trained to do in the situation? What about a stunt driver? I think regardless or morality the average person would slam on the brakes. so i go with slam on the brakes, i have yet to meet a person that when faced with killing a squirrel or avoiding anything in the road ever, decided to veer into a wall at 70mph. which would very likely end up bouncing off the wall and killing them anyways	Human driver will brake not swerve and hit wall	2	2.2
124.	This hypothetical situation is dumb. What are the parameters surrounding this "series of unfortunate events" that leads to a choice between killing a the occupant verses the crowd when: 1) The car can swerve/turn on a dime without rolling/flipping/drifting right into the crowd anyways; 2) The car can't break in time, assuming 60-0 break length of 115 ft and a 30-0 of 30 ft(is a self driving car programmed to speed, or does these kinds of cars get increased speed allowances?); 3) The car can't detect the traffic light or pedestrian pattern in time to stop normally, but successfully detects pedestrians in time to decide to swerve. Furthermore, why is a self-driving car driving at a speed where it can't break in time while the possibility exists for pedestrians to be legally on the road? Seems like that is why we generally limit speed	Unrealistic Circumstances are unknown	1	1.2

	on roads anyways. I would think the self-driving car would be the apex of lawful driving practices. Also, some other problems: 1) How does the car detect a crowd and decide to swerve when the possibility exists that crowds are not small packs and it could very well swerve into more people on the sidewalk in different speeds/areas of crossing? 2) Does the ethics change if people are in the street illegally (crossing a busy highway or street and not obeying traffic signs; illegally blocking traffic for social reasons, etc)? 3) Why do we assume that pedestrians should not be accountable for being in the street and not observing traffic around them before crossing streets? 4) What happens in the event that anarchist and criminals start jumping in front of cars to make them swerve and kill/injure their occupants? 5) Would a self-driving car not have standard and/or more robust safety features to protect occupants in the event of collisions (did we decide self-driving cars are 100% safe and thus remove airbags, seatbelts etc)? While I think ethical discussions surrounding machines and automation are good, I think this scenario is dumb.			
125.	True, that driver would instinctively in an dangerous situation save his life, car and occupants. Which unfortunately can be bad for people not in the car. And that also make pedestrians more careful with cars around. For reason of pedestrians not to lose being so careful with cars around, and owners of cars have trust in their cars, it is strongly recommended self-driving car to simulate driver's action in such situations: self-driving car's purpose in safety would be to make drive safer only that way, that the car would calculate and predict situations before, which can save passengers in situations where driver would not think of. As a result of using such cars statistically would be less accidents, which is an appropriate benefit of using them. Also car soft might make a mistake, e.g. thinking that 10 dolls, animals (or anything else what could be mixed with human) on the street are people, or made-up or accidental situation which can trick car's soft, therefore inevitably killing passengers for overthought reason, which would definitely not be acceptable. Car should be trustworthy by one's inside, and other's around should be careful with inappropriate actions towards them.	Human drivers will protect themselves Pedestrians must pay attention Such cars must simulate human drivers Systems could confuse between dolls and pedestrians and this is fail and unacceptable	1 3	1.4 3.3
126.	In our ultra connected world, maybe the pedestrians could be alerted before crossing the road... The cars will not be the only connected objets in the street. The humans too....	Such cars would warn pedestrians	2	2.4
127.	I would think that if speeds were sufficient that the car could not stop, it would also mean pedestrians were unexpected, perhaps j-walking or there was an accident. In either case, at those speeds, there is about to be a huge collision because freeways designed for self driving cars will have them spaced very close together, functioning as a unit. But if one of them did turn into the wall, it would cause the others to collide with it's flaming carcass and turn into a pileup, then the AI would shut down the freeway. Pedestrians have no business being on a road like that anyway. If they are in the city, they should cross at crosswalks, but the cars	If not avoidable situation, then pedestrians are jaywalkers Unrealistic	3	3.3

	should be travelling slowly enough that they could actually stop safely. The scenario is certainly exaggerated, but it seems to me they are after solutions to more finely detailed problems than the one described, which is ridiculous, but also a first step in understanding and developing a decision-making computer.			
128.	Ok, I can expect this from the French ;), but from smart people at MIT???? How come you did not raise issues with moral hazard of the choices offered? If cars are made to sacrifice car passengers for the sake of motorcyclists, bicyclists, pedestrians, etc. the "advantageous" groups will be behaving recklessly causing more deaths of innocent car passengers ...	This will lead to more deaths if such cars protect pedestrians over passengers	3 1	3.4 1.3
129.	I suspect initially at least, autonomous cars will only be fully autonomous on restricted access highways like interstates and on city streets and other non-restricted access roads will be manually driven or certainly less 'autonomously', again at least initially. This should provide a learning curve for manufacturers and consumers to develop answers to such ethical dilemmas. My view is the vehicles should be programmed to protect the occupants first, if nothing else because to cause scenarios like the ones illustrated, the pedestrians(or other vehicles in question will most likely be in violation of the rules that have been set forth for roads on which autonomous vehicles will be designed and programmed for. For instance, theoretically all autonomous vehicles will be in constant contact with other like vehicles in the general vicinity. Therefore these situations should rarely if ever occur unless someone or something is in violation of the preset programming, intentionally or unintentionally. And sadly, just as it is today, incidents like those described will be settled like they are now, by attorneys in a court of law in conjunction with insurance companies. Hopefully though autonomous vehicles will make such conflicts far less common.	Full self driving cars will be allowed only in highways Semi autonomous in cities Protect passengers first Pedestrians may break the law Such cars will contact each other	2 1 3 1	2.1 1.4 3.3 1.1
130.	I own a Volvo x90 2015 and i can tall ypu that than questions are never going to be asked. Why? If we analyze the case of the crowd, the car more than 200m away would already detect the obstacle. Another case tht i can tell you once happend to me is that i was driving at 60kph and the car suddenly stopped in a crosswalk and 1 second after a bike crossed in front of me, why? Becase the machine detected that a bike was coming and i wasn't slowing down. And i can five u lots and lots of examples. So technology will prevent this case of scenarios.	Such cars better than human drivers in detecting surrounding	1	1.1
131.	I think that there is a bit of a problem here, and I guess it not to far fetch. The premises is lets build cars that are as incompetent as humans and see if they make the same errors. With human desire to short-change everything that is entirely possible. It would more likely that the car can be programmed for a variety of scenarios with the least effect. The car can adjust speed, and direction faster than a human can. I think the question that the article poses is rather stupid. Do you think a human would drive head on into a wall on purpose? NO. They might swerve away from the crowd but still hit a few and probably bang the wall. driver lives,	Human drivers will not swerve and hit wall Unrealistic Such cars will be programmed for such scenarios	1	1.1
132.	what if the person on the sidewalk is einstein and the others are suicidals that want to die anyway?	what is pedestrians include doctors and criminals	1	1.5

133.	It's probably also worth considering that a given crash is more likely to kill a pedestrian than the occupant. Self-driving cars may be designed to higher passenger safety standards so that in the event they have to risk the well-being of the occupants they are more likely to survive without serious injury anyway. Look at high-speed F1 crashes - you're talking about 150mph+ speeds hitting a wall and there's often no serious injury. The first fully autonomous vehicles are liable to be operating at relatively low speeds (e.g. city car travelling 30mph), and there are very few scenarios that a passenger would be killed by a crash where the car has taken some action, compared to the risk to pedestrians at that same speed.	Such cars have safety features which will protect passengers Self driving cars will drive slow	1 1	1.3 1.1
134.	I would want it to be able to account for who is on board. An adult in the car will have a higher chance of surviving a crash than the pedestrians, but what if there is a baby in the car. If I were the driver, I would choose to sacrifice myself, but would not choose to sacrifice a baby, plus a baby would have a hard time surviving in a crash that there is a chance of an adult surviving.	Adults may survive from accidents but babies not	1	1.5
135.	Simple: eject button with parachute for driver and passengers, send email to insurance and go against wall to save crossing people	Eject passengers	2	2.2
136.	I think we need to have three options and the driver picks how they want their car set up, are you more conservative with passengers, the public, or meet somewhere in the middle and use a lowest probability of loss scenario. This eliminates the liability on the car companies as the driver was ultimately responsible for deciding how the car behaves in that situation the same as if he were driving. Also there is the potential you come off an exit on the highway into a city during a parade or something where a street that is normally open is full of people. While the car should still stop and the odds are high you'd be going slow in that scenario anyway due to traffic, inevitably a similar situation will play out.	Owner could choose car rule to hold responsibility	2	2.3
137.	Does anyone know how many times in their lives are drivers faced with such choice? Whatever the number may be, it is very close to zero.Of those extremely few who are faced with the dilemma, how many will be able to act effectively on their choice? Damn few, again. So we are talking here about a potential stumbling block to cover an infinitesimal probability. That, folks, is called paranoid waste. Meanwhile, a truly autonomous vehicle will react many times faster than any driver, with a level of unswerving attention to the road and traffic, on all sides, that any being with two eyes and biological constraints cannot even imagine. On radar, lidar and infrared, too, with LAN connections to road signs and other vehicles. The dilemma will never arise, for all practical purposes. The autonomous car will not be speeding, will not be distracted and will stop in time, without having to kill anyone. The clincher: are human drivers required to demonstrate their ability to pass this test, before they are given a licence? If they were, we would have very few accidents, with only a half-dozen drivers on the road, in the whole world. French Luddites, clutching at straws? I hope your silly jobs are the first to become redundant.	Unrealistic Self driving cars faster than human in reaction Such cars will drive slow	1	1.1
138.	No one would buy a car that is programmed to kill them. if killing for the "greater good" is justifiable, then taking out politicians or vigilante justice against criminals who avoid the justice system over a technicality would be "justifiable". The issue with the greater good	Will not buy killer car Utilitarianism is not good for individuals	3 1	3.1 1.5

	argument and with much utilitarian theory is that it doesn't take into account the individual which is an issue, because individuals make up the many. While it may seem better or more moral to save ten and sacrifice one this is wrong because it devalues the individual and since we are all individuals it is the individual that is more important than the whole in the microcosm of a car accident. Any self-driving car must always defer to the safety of the driver in the first instance, protecting the individual and ultimately serving the greater good by protecting all individuals. Not ten or twenty but the millions who exist on the planet today. In acknowledging the value of the individual we protect the whole lest we all become lambs to be sacrificed on the altar of artificial intelligence as it sees fit.	Such cars must protect individuals		
139.	The car can be programmed to kick the driver out and then swerve to avoid hitting the other people.	Eject passengers	2	2.2
140.	Did anyone consider using a horn? The speed of sound at sea level is 343 meters per second or 768 mph, unless you are traveling close to this speed,it would be unreasonable to swerve or run over the pedestrians because the pedestrians will hear the horn and know to get out of the way. Depending on the amount of people, an eject button for yourself and the passengers would also be reasonable while the car swerves to hit a wall.	Warn pedestrians Eject passengers	2 2	2.4 2.2
141.	The key is the choice of words. An accident is avoidable and as a result there is no need to create an algorithm to minimise death as there should never be an accident, they would always be avoided. An incident however may result in death so the the key is to reduce the chances of incidents. This is simple as the algorithm should simply minimise the factors that contribute to incidents. The main factor in incidents is speed and while speed does not necessarily kill, it is the sudden exchange of energy when two thing meets as a result of high speeds. Ergo take the issue of speed out of the equation (or reduce it effects on the equation) and the problem becomes very clear. We do not need self driving cars to create a solution to the current problem, we simply need to stop the current problem, people and their selfishness that their right to be somewhere sooner overrides other peoples rights to modal safely.	Such cars always avoid accidents No need for self driving cars to solve problems	1 3	1.1 3.2
142.	There are many ways to avoid this problem. We are an advanced society, surely we can come up with an answer. Everyone's ideas accounts for something. Of course vehicles aren't engineered to come to a complete halt-if it were in a situation described in the article-but there are many different ways to get a vehicle to stop that won't result in injuring the passengers or anybody else involved.	Engineers will solve the problem	3	3.5
143.	This really seems to be a false choice or false dichotomy, on several counts. - Before this becomes a "choice of who to kill", we need to build sensors that can rapidly, in a time scaled by the speed and stopping distance, identify the presence and number of people in the path of the vehicles, their position and direction of travel. The vehicle would also need to identify that the shoulder was unoccupied, not residential homes, parks, storefronts, or other likely occupied space. - The question posed is should the vehicle choose to kill passengers or folks in the road. What about stopping, evasive manoeuvres in the roadway, or both; better, I would expect early warning from the roadway itself, allowing the car time and space to slow or re-route well before the jay-walkers. This is a technical problem, not a moral one. - As this board demonstrates - *people* can agree on the right	Unrealistic Before that we need systems to prevent that Warn and alert jaywalkers It is technical problem not ethical	1 2	1.2 2.4

	course. The right answer, in the absence of more or earlier information, is probably brake, steer to the biggest gap and stop as quickly as possible, *in the roadway*. Drivers are expected to do their best, but jay-walkers take their chances with human drivers, why would we necessarily expect that the risk is lower with automated systems?			
144.	The ethical debate here with self-driving cars isn't much different than the dilemma faced by manufacturers of war drones that are equipped with enough programmed artificial intelligence to recognize an enemy target and take action without the drone operator being involved in the decision. If you let a machine do the reasoning, then you must be prepared for the result.	If you want machines to reasoning, prepare for the outcome	3	3.5
145.	I would want a setting to either minimize loss of life or protect occupants as the highest priority. Thus the car could be directed to best reflect the ethical choice of its passengers.	Protect passengers	1	1.4
146.	This dilemma is old and studied by philosophers for centuries. He appears in many books of ethical dilemmas, including Julian Baggini, "The Pig Philosopher". It is an ethical problem with moral boundary. There is no way to create an ethical or moral algorithm. Always miss something, whether as a result of the law (moral) or on the basis of principles (ethics). There is a case in moral studies important to start that conversation. The case of Phineas Gage (1822/1861) which was an American worker who, in an accident with explosives, had his brain pierced by a metal bar, surviving despite the severity of the accident. After the incident, Phineas, who apparently had no sequelae, showed a marked change of moral behavior, and object to case studies well known among neuroscientists, proving that moral judgment is included (physically) somewhere in the brain. In the above case, my opinion is that either the driver will buy a car with choices already defined (kill / no kill) and still be liable for the act (ethical judgment a priori), or the driver will hear and decide the time and also It will be held responsible for the act (subsequent trial). In my view this decision involves the ethical and moral living experience that the driver had a lifetime. If this experience is intense, he knows which way and responsibility he has to follow. If there was no intense experience that sense, it does not matter to him if the car decides, for his view of ethics and smaller. That said, we must understand that we are not ethical all the time, all the time. In fact "we are ethical" in parts of the time of our lives, taking into account the ethics of good and evil is universal. This dilemma already happens when they die at this very moment, 11 people per minute from hunger.	No way to build ethical algorithms Owner either buy programmed car to kill or not or decide We are not ethical all the time	3 2 3	3.2 2.3 3.5
147.	I think this push toward driverless technology is moving forward too fast. The companies developing the technology are attempting to push the cars onto the market before either the technology or society is ready for its deployment. It has taken more than 50 years for robotics to advance to the point that driverless vehicles are even remotely feasible, but the companies pushing this tech are acting like these can all be rolled out either now or within just a few more years. I believe the latter is way too premature since they are just now starting to grope with the *real* problems for this technology: how to make driverless cars interact safely with other human drivers. I think this is a sufficiently	Companies rush to market before society ready Companies see their profits not risks to human lives	3	3.5

	challenging problem that it may easily take several more decades if even feasible. Let's not let an excessive enthusiasm for gadgets and profits lead us to rush out a technology that results in a lot of needless human deaths because it wasn't ready and was poorly thought out. Having worked in tech companies before, I can assure you that the management of these companies have not done a very good job of listening to dissenting opinions either within or outside of their organizations. Potential profits are mostly what they see, not risks to human life.			
148.	This has been going around for a while but my perspective on this situation is that self-driving cars shouldn't sacrifice their passengers because if they never break the law than the person that would be hit by the car would be the one at fault.	Shouldn't sacrifice passengers because of jaywalkers	1 3	1.4 3.3
149.	The eventual solution must also deal with two issues that the article does not mention: uncertainty, and misconduct. Maybe that is a pedestrian in the road, and maybe it's a tumbleweed, or something that fell off a truck. And maybe it's a person-sized balloon that somebody dropped off the overpass, just for the fun of it. I vote for protecting the occupants.	Protect passengers There are uncertainty and misconduct	1 1	1.4 1.2
150.	Something I've been thinking about is the legality of autonomous car crashes. These companies will undoubtedly make us all sign wavers that say that we are responsible for all accidents and not them. However, we are not the ones driving the car. This brings up a number of important questions. 1. Who will pay for car insurance? Given historical practice I'm assuming the driver. 2. If an autonomous car is found at fault for an accident, does the driver's insurance go up? They were not even driving at the time of the accident. Does everyone using that company's software have an increase in insurance since the software was driving? 3. When are person injured by an autonomous car wants to sue, who do they sue? 4. Is it possible that the manufactures of these autonomous cars will also have a group insurance plan for its customers to get around these legal issues?	We may sign waivers for responsibility	2	2.3
151.	I think self driven cars still have many issue an there are many consequences of using it. At some extent I like technology until and unless it won't harm anyone. But this self driven car will	Self driving cars are not yet ready	3	3.1
152.	I have a different answer: External crash bags. If the car detects it is going to crash into a person it inflates crash bags around the bumper to minimise the impact. Patent pending, I expect to make a dollar fifty out of this idea :)	Airbags outside the car	2	2.2
153.	There is a huge flaw in the main initial idea itself. Why need to force the device to choose only either to protect the driver or the pedestrian, especially there are so much more to consider in a real dynamic circumstances? This is like the concept is putting huge bias into the root of machine before any real possible encounter. Why cant the car's intelligence protect both or minimize (if not zero) the total fatalities irrespective of whether the driver or pedestrian? It should be able to take account a large angle of inputs from the surrounding for calculation to reduce all the possible damage (while considering faulter) to all sides of parties given lives will be the top priority. In that given split of second, human counts of inconsistent, inaccurate, emotional and bias reflexes to make a prompt decision.	Why we force machines to decide during dynamic environment Why we let machines to minimize damages Lives is priority Machines are better than us and we don't want to program them to be like us	1 3 1	1.1 3.2 1.2

	Therefore, machine can do better than that in term of speed of input calculations, accuracy and consistency to make better decision for minimal fatality irrespective of whether user or pedestrian. The earth is enough of greed and selfishness by human, we do not need to program such traits in machine to add more greed and selfishness to the earth.			
154.	What if instead we build bicycle-friendly cities, and everybody rides a bike and everybody is happy (and alive)? And then we don´t have to decide who should die has nobody dies...	Build friendly cities and no body will die	2	2.1
155.	What if there were 100 people on the other side of the wall and the wall was made of glass? The car isn't that smart.	What if there are more people behind the wall	1	1.2
156.	What if someone, say is attempting suicide by walking in front of a large vehicle on purpose, and despite their intentions, they end up indirectly committing murder? Hmm. Also, in these examples the author assumes 10 people will die. Does the electronic driver really know that 10 people will die? does it really know for sure that the driver will die? if a driver is moving that fast to kill 10 people, how often are their walls right there so close that will kill the driver if they run into them? Do self-driving cars not have airbags?	What if someone jump in front of the car Crashes outcomes are unpredictable	3 1	3.4 1.2
157.	I think it should definitely protect the passengers and it should work according to the traffic regulations applicable in the country you are "driving" your driver-less car. Think about it, if you're about to have an accident in a manual car, in that moment you certainly don't have any ethical thoughts. Neither the car should have Considering your example: 10 people crossing the street; the car has a too high speed and cannot stop in time Here I see two possibilities: 1. The people are crossing the street on a crosswalk. In this case, the car must not have in any circumstance a speed too high not to stop in time. It should adapt the speed according to the traffic indicators, weather, etc. 2. The people are not crossing on a crosswalk (they are on a road/highway). Basically they are not allowed to be there. They should be aware they can be harmed and they endanger also other people. The car should, of course, try to avoid them, overpass them, etc. But the car, for sure, must not crash and maybe kill it's passengers. Everyone must be self aware of the dangers that arise when one crosses in a forbidden zone. The car must always protect it's passengers! Let's try not to become stupider than a car and we can be able to survive.	Protect passengers Human drivers in accident will act without ethics Every one must be self aware	1	1.4
158.	These scenarios aren't very realistic. Provided the occupants of the car were wearing restraints and all air bags were functional, it is highly unlikely hitting a wall would result in the occupants' death. The velocity needed to create such fatal force with modern protections would require the vehicle to be traveling at highway speeds. In which case there is little possibility of encountering pedestrians. Unless they're escaped convicts, which would create a far more interesting moral dilemma.	Unrealistic Cars are safe	1	1.3
159.	A lot of articles like this begin with "assume you know for certain the following facts".. e.g. you have 2 choices, either make a change that will kill 1 person or do not change and kill 10 people. In reality all the data will be so fuzzy and things will be moving so fast that you (or a computer) is going to be able to assign a very low	Unrealistic In reality the choices are fuzzy Such cars will stop without swerving	1 1 2	1.2 1.1 2.4

	confidence to any of these preconditions. "What if the 10 people are young and agile and can jump out of the way"? "What if the 1 person has terminal cancer?" "What if, when you try to swerve to hit the 1 person, your car rolls and hits the 10 people anyway?".. An finite regression of "what ifs". Realistically the computer is going to have to pick the simplest solution - hit the breaks to try and slow down as fast as possible without skidding, stay in your lane so that people can predict your path, sound the horn, maybe deploy airbags? Anything more fancy would be likely to have unintended consequences.	Pedestrians will know Alert and warn them		
160.	How about we simply have a switch with two (or more) modes in each car and let the driver decide which one to select. One mode could be to maximise the safety of the occupant and another mode could be to minimise overall harm. This way it puts the ethical burden onto the occupant, and they will have to consider and live with the consequences for the rest of their lives.	Owner could choose between modes in cars	2	2.3